INTERSPEECH 2014

CELEBRATING THE DIVERSITY OF SPOKEN LANGUAGES

14-18 SEPTEMBER 2014
SINGAPORE
MAX ATRIA@SINGAPORE EXPO

HTTP://WWW.INTERSPEECH2014.ORG
Organizers

Diamond Sponsors

Gold Sponsors

Silver Sponsors

Bronze Sponsors
ISCA is a non-profit organization. Its original statutes were deposited on February 23rd at the Prefecture of Grenoble, in France by René CARRÉ and registered on March 27th, 1988.

The association started as ESCA (European Speech Communication Association) and, since its foundation, has been steadily expanding and consolidating its activities; it has offered an increasing range of services and benefits to its members and it has put its financial and administrative functions on a firm professional footing. Indeed, over the ten years of its existence, ESCA has evolved from a small EEC-supported European organisation to a fully-independent and self-supporting international association.

At the General Assembly that took place during the last Eurospeech conference in Budapest (September 1999), ESCA became a truly international association in the global field of speech science and technology, changing its name to ISCA (International Speech Communication Association) and modifying its statutes accordingly.

The purpose of the association is to promote, in an international world-wide context, activities and exchanges in all fields related to speech communication science and technology. The association is aimed at all persons and institutions interested in fundamental research and technological development that aims at describing, explaining and reproducing the various aspects of human communication by speech, that is, without assuming this enumeration to be exhaustive, phonetics, linguistics, computer speech recognition and synthesis, speech compression, speaker recognition, aids to medical diagnosis of voice pathologies.
The Chinese and Oriental Languages Information Processing Society, or COLIPS in short, is a non-profit professional organisation that was established in 1988 to advance the research of Chinese and other Asian languages. It promotes the free exchange of information about information processing of these languages in the best scientific and professional tradition.

COLIPS organizes international conferences, short courses and seminars for members and the public. COLIPS publishes the International Journal of Asian Languages Processing four times a year that is circulated world-wide. Having its members from all over the world, COLIPS is based in Singapore. It is one of the founding members of Asian Federation of Natural Language Processing (AFNLP).

COLIPS proudly co-organises INTERSPEECH 2014 with ISCA!
Singapore’s largest ICT research institute, I²R (pronounced as i-squared-r) is a member of the Agency for Science, Technology and Research (A*STAR) family.

Established in 2002, our vision is to power a vibrant and strong infocomm ecosystem in Singapore. I²R focuses on conducting mission oriented research to address key challenges faced locally. At I²R, intelligence, communications and media (ICM) form our 3 strategic thrusts.

Our research capabilities are in information technology, wireless and optical communication networks, interactive and digital media, sensors, signal processing and computing. We perform R&D in ICM technologies to develop holistic solutions across the ICM value chain and we believe that the greatest impact is created when research outcomes are translated into technologies our partners can readily deploy at a competitive advantage.

The Human Language Technology Department at I²R proudly co-organises INTERSPEECH 2014 with ISCA!
WeChat is an IM mobile app developed by Tencent Inc. providing multimedia communication with Text Messaging, Voice Messaging, Video Call, Group Chat, Emoticons and social functions like Moments, Look Around, Shake, Game Center. WeChat Official Account Platform helps enterprises connect target customers. WeChat was released in January 2011 and acquired 390 million monthly active users in early 2014. It's available on iOS, Android and other mainstream mobile platform supporting 19 languages.

Temasek Laboratories@NTU (TL@NTU) is a partnership established by the Defence Science & Technology Agency and Nanyang Technological University (NTU) in 2003. Our mission is to explore the frontiers in science and develop strategic technology that will deliver effective solutions to enhance the defence and security of Singapore. We carry out innovative R&D programmes that lead to technological breakthrough in defence science and technology. Our researchers also facilitate the transition of technology to applications and provide technical advice to our partners.
WeChat

The new way to connect

Scan to start your WeChat life.

600 Million Users

Instant Message  Social media  LBS  Public platform

http://www.wechat.com/en
Our founders applied their research to an interesting problem and invented the world’s most popular search engine. The same spirit holds true at Google today. The mission of research at Google is to deliver cutting-edge innovation that improves Google products and enriches the lives of all who use them. We publish innovation through industry standards, and our researchers are often helping to define not just today’s products but also tomorrow’s.

Baidu was founded in 2000 by Internet pioneer Robin Li, creator of visionary search technology Hyperlink Analysis, with the mission of providing the best way for people to find what they’re looking for. Our deep understanding of Chinese language and culture is central to our success and this kind of knowledge allows us to tailor our search technology for our users’ needs.

Amazon strives to be the most customer-centric company on the planet. We’re building speech and language solutions which will enhance user interactions with Amazon products and services. Products such as Fire TV and Dash are illustrative of the user-delighting spoken language understanding solutions Amazon is building. We’re working hard, having fun, and making history.

Gold Sponsors

Amazon

Baidu

Google
Microsoft is all-in on NUI so our customers can do more everywhere. Our scientists and engineers in dialogue, speech, language, machine learning and synthesis develop and deliver robust, natural and scalable spoken language systems across a rich set of scenarios, languages and devices. We look forward to meeting you at INTERSPEECH in Singapore.

Founded in 2000, SinoVoice a high-tech enterprise focusing on researching Human-Computer Interaction technology, is the best speech technology supplier in China. In 2011, SinoVoice launched HCICloud platform (www.hcicloud.com) providing complete HCI service like TTS, ASR, VPR, NLU, HWR, OCR, MT, etc. around the world.
Unisound, a professional voice interactive service provider focusing on automatic speech recognition and language processing technologies, has obtained several independent intellectual property rights of the world’s top intelligent voice recognition technology. Founded on June 29, 2012, the company is now headquartered in Beijing, with a branch in Shanghai and an office in Shenzhen.

Datatang is an innovative company focusing on big data globally. Datatang operates China’s largest data exchange platform and is engaged in collecting, annotating, and customizing data based on clients’ needs. By assisting clients to waive trivial yet necessary data cleansing procedure, Datatang makes approach to the highest-value data in a more efficient way.

iFLYTEK Co., Ltd. is a listed software enterprise founded in 1999, which is dedicated to the research of Multi-lingual speech technologies, development of software and chip products, provision of Multi-lingual speech information services, and integration of E-government systems. The core technology of iFLYTEK, represents the top level in the world.

Nuance Communications, Inc. is a leading provider of voice and speech recognition solutions for businesses and consumers around the world. We design and deliver intuitive speech driven technologies that help people live and work more intelligently through our speech/machine interactions. We provide the tools to inform, to connect, and to empower people to be more productive and creative.

Established in 2008, Vcyber Technology Co., Ltd., headquartered in Shenzhen, specializes in developing in-vehicles human machine interaction systems and telematics platform operation. Vcyber is committed to creating a safe and convenient in-vehicle experience, and is capable of offering complete telematics solutions through integration of vehicle data, smart terminals and backend service resources.
Bronze Sponsors

ELRA

ELRA, the European Language Resources Association founded in 1995, is a not-for-profit organisation whose mission is to collect, distribute, and validate language resources (LRs) and tools for research institutions and companies. ELRA, through the agency ELDA, also produces LRs, and promotes the LRs and evaluation field. The LREC conference organised by ELRA is the main event of this field.

Facebook

Founded in 2004, Facebook’s mission is to give people the power to share and make the world more open and connected. People use Facebook to stay connected with friends and family, to discover what’s going on in the world, and to share and express what matters to them.

IBM Research

A main focus at IBM is that the world is our laboratory, and no matter where discovery takes place, IBM researchers push the boundaries of science, technology and business to make the world work better. With this goal in mind, the Speech and Language algorithms group at IBM tackles problems in speech recognition, dialogue systems, natural language processing and machine learning. IBM Research is a global community of forward-thinkers working towards a common goal: progress.

PCS Security

PCS Security is a Systems Integrator specialising in Infocomm (ICT) and Homeland Security. We work with cutting edge technologies to deliver effective and secure solutions for our clients to fulfil security and operational needs. Our vision is to be a leader, preferred partner and the trusted vendor for ICT in Homeland Security.

Raytheon BBN Technologies

The Speech, Language and MultiMedia department of Raytheon BBN Technologies (BBN) comprises more than 150 people dedicated to innovations in research, development, and product offerings. Current research activities include applying machine learning algorithms to speech processing, language understanding, machine translation, human-machine interaction, image/video processing, and forecasting from streaming data.

Speechocean

As a language resource and resource-related services provider with a 15-year history, Speechocean is capable of providing its customers one-stop linguistic solutions in the field of HCI/HLT in 110 languages and accents. Speechocean currently has over 500 large databases such as speech, text, lexicon, image, etc., available to its commercial and academic customers in creating diversified values.
Samsung Electronics leads the global market in high-tech electronics manufacturing and digital media. SRC-Beijing (Samsung R&D Institute China - Beijing) is one of Samsung’s overseas R&D centers, located in Beijing, China. SRC-Beijing focuses on the research & development of next generation communication, advanced multimedia, big data, 3D display, mobile terminal and other emerging technologies. We foster innovative ideas that advance technology, produce new products and improve the everyday lives of our customers.
Media Partner

Exhibitors

Amazon
Appen
Baidu
Datatang
ELRA
Google
iFLYTEK Co., Ltd.
INTERSPEECH 2015
INTERSPEECH 2016

ISCA
Linguistic Data Consortium
Northern Digital Inc.
Nuance
Samsung
Speechocean
Springer
VoiceBox
WeChat, Tencent

Supporters

Yandex

YourSingapore.com
## Contents

Welcome to INTERSPEECH 2014  ................................................................. 16

Welcome to Singapore ................................................................. 20

INTERSPEECH 2014 General Information ............................................. 22

Social Program ................................................................. 25

Organizers ................................................................. 26

Technical Program Committee .................................................. 28

Scientific Review Committee ...................................................... 30

Future INTERSPEECH Conferences .................................................. 35

Satellite Events ................................................................. 36

Keynote Speeches ................................................................. 37

Special Sessions ................................................................. 42

Special Events ................................................................. 45

Tutorials ................................................................. 46

Awards ................................................................. 51

Daily Schedule ................................................................. 54

Session Index ................................................................. 59

Abstracts ................................................................. 61

Author Index ................................................................. 187
Welcome to INTERSPEECH 2014

Message from the General Chair and Co-Chair

Welcome to Singapore - where INTERSPEECH makes a return to Asia!

INTERSPEECH 2014 marks the 15th annual conference of the International Speech Communication Association (ISCA). We are indeed honored to pen the first words of the INTERSPEECH 2014 proceedings. In the past year, research in speech communication has continued to thrive across Asia and all over the world. On this occasion, we share with you the excitement of our community as we gather again at our annual conference in Singapore.

Among the 7,105 living languages and dialects that Ethnologue documented in 2013, 2,304 are spoken by Asians, representing more than half of the world’s population. Unfortunately, only less than 1% of them have been actively studied by the scientific community. INTERSPEECH 2014 embraces the theme ‘celebrating the diversity of spoken languages’ to call for an attention to extending speech science and technologies to a wider variety of languages and to benefit the whole of humanity.

There is no better place to celebrate the diversity of spoken languages than Singapore. One of the most remarkable aspects of Singapore is the truly cosmopolitan nature of her population. The major ethnic groups in the country include the Chinese, Malays, Indians, and Eurasians. As a reflection of its collage of cultures, Singapore has adopted four official languages in its constitution, namely English, Mandarin Chinese, Malay and Tamil. Almost everyone in Singapore speaks more than one language, with some speaking as many as three or four in their daily life.

INTERSPEECH is made possible by the dedication and hard work of our community. We are indebted to many. The INTERSPEECH track record of excellence continues this year. This would not have been possible without the advice and support of the ISCA Board, INTERSPEECH 2014 Advisory Committee, and previous conference chairs. Special mention must be made of the fact that INTERSPEECH 2014 is truly international. It is organized by a core team from Hong Kong and Singapore, and an organizing committee with members from 16 economies. The diversity of our people is one of our core strengths that adds tremendous value to conference organization. We thank all of them for volunteering their time and energy in service to our community.

In particular, we would like to express our gratitude to the Technical Program Chair Professor Helen Meng and Co-Chair Dr. Bin Ma for their leadership in putting together an exciting and high quality technical program, the Local Arrangement Chair Ms. Swee Lan See who did a superb job coordinating local logistics. Special thanks to the Workshop Chair Dr. Chai Wutiwiwatchai, Tutorial Chair Professor Eliathamby Ambikairajah, Plenary Session Chair Professor Brian Mak, Special Session Chair Dr. Tomi Kinnunen, and Technology Show Chair Dr. Alvina Goh for chairing the respective committees. Thanks are also due to the Publication Chairs, Professor Eng Siong Chng and Professor Lei Xie, the Exhibition Chair Dr. Xiong Xiao, Publicity Chair Dr. Anthony Larcher, Sponsorship Chair Dr. Nancy Chen, and the Webmaster Dr. Lei Wang for their dedication and tireless efforts. Finally, we are most grateful for the generous sponsorship from the industry, and the overwhelming supports from two local organizers, the Chinese and Oriental Languages Information Processing Society and the Institute for Infocomm Research, A*STAR, Singapore.
the ISCA Archive during the conference. With the strong support from Sebastian Möller and his team, we updated the mobile InterspeechApp application to provide the latest technical and social programs at your finger tips. The revamped InterspeechApp also features a natural language interface and a personalized scheduling service. During the week, Twitter (#interspeech2014), ISCA’s Facebook and INTERSPEECH’s LinkedIn group will bring you the latest news and practical details.

Finally, it is our great pleasure to welcome you to Singapore and INTERSPEECH 2014. We hope that you will have an enjoyable and productive time in Singapore, and leave with fond memories of the multilingual experience at INTERSPEECH 2014. With our best wishes for a successful conference!

Haizhou Li  
Institute for Infocomm Research, A*STAR, Singapore  
General Chair

Pak-Chung Ching  
The Chinese University of Hong Kong, Hong Kong SAR, China  
General Co-Chair
On behalf of the INTERSPEECH 2014 Technical Program Committee (TPC), we warmly welcome you to the 15th Annual Conference of the International Speech Communication Association (ISCA) from 14 September to 18 September in Singapore. We are honored to have the opportunity to undertake the challenging task of working out the Technical Program, which required close teamwork and a strict timeline. We have benefitted greatly from the advice and experiences of the leaders of previous years of INTERSPEECH, especially INTERSPEECH 2013. We carefully laid down the technical review process for INTERSPEECH 2014 to strive towards decision-making based on technical merit, and followed through with it faithfully. We present a succinct description of the process below, which we believe will be a useful reference for the future.

The technical review process of INTERSPEECH 2014 began with the definition of 12 technical areas that underpins the scope of our program. This led to our team’s formation with 36 Area Chairs who have kindly accepted our invitation to serve the conference. We also worked with our Special Session Chair in issuing a call for Special Sessions, leading to 12 provisionally accepted special sessions. The previous practice of INTERSPEECH 2013 was followed and the START v2 paper review system was adopted. Each Area or Special Session was set up as a track in the START system, and was managed by the respective Area Chairs and Special Session Chair. We invited 1,245 reviewers of whom 914 kindly agreed to serve.

INTERSPEECH 2014 received 1,173 complete submissions, of which 105 targeted the Special Sessions, and all were sent out for review. We leveraged the automatic reviewer assignment algorithm in START that performed matches between the reviewers’ declared areas of interests and each paper’s declared topics. The algorithm attempts to perform an exact match between the preferred areas of each reviewer with the top-ranking topic of each paper. Failing that, the algorithm attempts a backoff match with the extended list of areas. Attention was paid towards avoiding conflicts of interests (CoI), drawing upon information such as the declared affiliations and email addresses of the authors and reviewers. Hence it is worthwhile to note that generic email accounts are less preferred in the START system. The automatic assignments and CoI were further reviewed manually and refined by the Area Chairs. For papers submitted to Special Sessions, two reviewers were selected by the Special Session proposers and one reviewer was assigned by the automatic reviewer assignment algorithm – this arrangement was designed with due regard to the nature of the special and emerging topics of the submissions.

Our team tried hard to ensure that each submission could receive at least 3 reviews. This was not an easy task as a certain number of review assignments were not received before the deadline. Our Area Chairs enlisted a number of crash reviewers for each area, who generously volunteered their time and effort to provide reviews at short notice. With these measures, we fulfilled the goal that there are no missing reviews for all the submissions.

The scores and comments uploaded by our reviewers for each submission were carefully considered by our Area Chairs. Issues such as the level of confidence of the reviewers, variations and inconsistencies in scoring, appropriateness of the review comments, discrepancies among reviews, etc. were all taken into consideration. Suggested accept/reject decisions from the Area Chairs were further discussed openly in a two-day, face-to-face TPC meeting held in Singapore between 29-30 May, 2013. The team consisted of 23 committee members, including the Technical Program Chairs, the Special Session Chair, at least one representative from each team of Area Chairs across all 12 Areas, the General Conference Chair and the ISCA President. All papers flagged by the Area Chairs were discussed individually by carefully examining all input scores and comments and meta-reviews were conducted wherever necessary. This decision process led to 614 accepted submissions – corresponding to an acceptance rate of around 52% which is consistent with both regular and Special Session submissions.
Organization of the accepted papers into oral and poster sessions was also discussed in the face-to-face meeting. While we considered that all 12 provisionally accepted Special Session proposals were good, we could only accept 6 with a sufficient number of papers to occupy the (limited) number of Special Session slots in the Technical Program. In addition, the Technical Program Chairs and Technology Show Chair also reviewed and accepted 22 Show & Tell proposals for demonstrations at the conference, and each is accompanied with a two-page paper. Overall, the INTERSPEECH 2014 Technical Program is structured into 42 oral sessions, 30 poster sessions, 6 special sessions and 2 Show & Tell sessions.

In addition, the INTERSPEECH 2014 Technical Program also features 5 keynote speeches, 8 tutorials and 6 satellite workshops. The ISCA Medalist, Dr. Anne Cutler, a pioneer in psycholinguistics, will give insights on learning about speech; Professor K. J. Ray Liu will deliver a keynote speech about decision learning in data science; Dr. Lori Lamel will talk about language diversity; Professor William Shi-Yuan Wang will talk about sound patterns in language; and Dr. Li Deng will share the achievements and challenges of deep learning. The 8 tutorials, which were selected from some 15 tutorial proposals, will be arranged into two half-day tutorial sessions on 14 September. The 6 workshops are the satellite events of INTERSPEECH 2014 and will be held either before or after the main conference.

Organizing the Technical Program of INTERSPEECH involves a huge amount of work which would not have been possible without the help and support of a large group of highly dedicated colleagues. We wish to thank our ISCA President, Professor Tanja Schultz, the ISCA Board, the INTERSPEECH 2014 General Conference Chair, Professor Haizhou Li and the entire Local Arrangements Committee for their support. We are most grateful to our Area Chairs who have offered tremendous technical leadership over the past year, promptly responding to all our requests even on weekends. Our appreciation goes to all the authors who have submitted high-quality papers, the Special Session proposers for submitting novel ideas on emerging research topics, and all the reviewers who have diligently assessed all the submissions. In particular, special thanks go to our crash reviewers, many of whom also served as our regular reviewers. We also appreciate the hard work of our Special Session Chair Dr. Tomi Kinnunen, Plenary Session Chair Professor Brian Mak, Tutorial Chair Professor Eliathamby Ambikarajah, Workshop Chair Dr. Chai Wutiwiwatchai and Technology Show Chair Dr. Alvina Goh. Last but not least, our gratitude goes to Professor Isabel Trancoso for always offering a helping hand and the wonderful Technical Program Chairs of INTERSPEECH 2013, Dr. Lori Lamel and Professor Pascal Perrier, who have continuously offered valuable advice and suggestions, as well as answering our numerous questions as we followed in their footsteps to take on new challenges.

We sincerely hope that you will find the Technical Program of INTERSPEECH 2014 intellectually stimulating and wish you a very enjoyable and productive time in Singapore!

Best Wishes,

Helen Meng
The Chinese University of Hong Kong, Hong Kong SAR, China
Technical Program Chair

Bin Ma
Institute for Infocomm Research, A*STAR, Singapore
Technical Program Co-Chair
Welcome to Singapore

Located at the southern tip of the Malay Peninsula, only 137 km north of the equator, the tiny “city-state” of Singapore is easily reached through Changi Airport, a major air hub in Asia connected to 270 cities in about 60 countries and territories worldwide. Singapore has one of the most extensive and efficient public transportation systems in the world. Travelling in the city is typically a quick and affordable affair.

A Brief History...

The history of Singapore started when the first settlements were established in the 13th century AD. Along the years, Singapore was part of different kingdoms and sultanas until the 19th century, when modern Singapore was founded under the impulsion of the British Empire. In 1819, Sir Thomas Stamford Raffles landed in Singapore and established a treaty with the local rulers to develop a new trading station. From this date, the importance of Singapore continuously grew under the influence of Sir Raffles. Singapore remained under British administration until the Second World War and became a Crown Colony after the end of the conflict. A brief period followed during which Singapore was part of the Federation of Malaya, before becoming independent in 1965 and part of the Commonwealth of Nations. Today, Singapore is a bustling cosmopolitan city of high-rise buildings, landscaped gardens and heritage-rich precincts. Home to a harmonious blend of culture, cuisine, arts and architecture, Singapore is rich in contrast and colour. In fact, you will find that Singapore embodies aspects of both East and West in many interesting ways.

A Multicultural City

Drawn by the lure of better prospects, Singapore’s immigrants brought with them their own culture, languages, customs and festivals. Intermarriage and integration helped knit these diverse influences into the fabric of Singapore’s multi-faceted society, giving it a vibrant and diverse cultural heritage. By the end of the 19th century, Singapore had become one of the most cosmopolitan cities in Asia, with major ethnic groups in the country being the Chinese, Malays, Indians, and Eurasians. Today, the ethnic Chinese form 74.2% of the Singaporean population, with the country’s original inhabitants, the Malays, comprising 13.3%. The Indians make up 9.2%, and Eurasians and Asians of different origins making up a combined 3.3%. Singapore is also home to many expatriates coming from countries as diverse as North America, Australia, Europe, China, Japan and India.

A Multilingual Environment

In Singapore's constitution, “Malay, Mandarin, Tamil and English” are the “four official languages”. In practice, Malay, Mandarin and Tamil are treated equally as they each correspond to one of the main ethnic groups present in the country. Inherited from the colonial history, English, has a special place in the Singaporean society as the only vector of education and the only official language that is ethnically neutral. No doubt that you will also appreciate the diversity of Singapore through the widely spoken and colorful Singlish (Singapore English) which mixes English, Malay, Hokkien, Teochew, Cantonese and Tamil.

The diversity of languages spoken in Singapore is not limited to the four official languages. Each of the three major ethnic groups has a number of spoken languages. Chinese Singaporeans, whose origins are predominantly from the south of China, speak Hokkien, Teochew, Cantonese, Hakka, Hainanese, Foochow, Henghua, Shanghaiese or Hokchia. Malay Singaporeans speak Malay, Javanese or Boyanese, while Indian languages in Singapore include Hindi, Panjabi, Gujarati, Tamil, Malayalam, Telugu or Kannada. Almost everyone in Singapore speaks more than one language, with some speaking as many as three or four. Most children grow up bilingual from infancy, learning other languages as they become older. With the majority of the literate population bilingual, English and Mandarin are the most commonly used languages in daily life.
… Located in a Garden

The “city-state” of Singapore hosts one of the two largest urban primary rainforests in the world and more than ten thousand species of plants. Nature lovers can explore Singapore’s natural charm via her islandwide park connectors that traverse Singapore’s most popular beach parks and rustic enclaves. Alternatively, you can explore the Bukit Timah Nature Reserve, Singapore’s oldest rainforest, or the Sungei Buloh Wetland Reserve, where you can learn more about our mangroves and spot mudskippers, crabs and other marine life. Skyrise greenery is another example of the clean and green Singaporean lifestyle. Roof-top and vertical greenery increasingly adorn Singapore's buildings such as the National Library Building and upcoming projects.

An International Economic Hub

Although small in size, Singapore is a bustling cosmopolitan city of high-rise buildings, landscape gardens and heritage-rich precincts. Her historically strategic location has enabled her to become a central sea port along major shipping routes. Singapore has one of the most open economies for international trade and investment in the world and a thriving financial centre of international repute. With more than 6,000 meetings, incentive travels, conventions and exhibitions held each year, Singapore is a major conventions and exhibitions hub of Asia.

World Class Attractions and Shopping

Even Singaporeans who have lived most of their lives here constantly discover something new. Singapore’s rich cultural artefacts, interesting historical monuments, idyllic nature spots and fun tropical spaces make her an enthralling leisure destination. In fact, Singapore is home to many firsts - the world’s first F1 Grand Prix night race, the sole Universal Studios in Southeast Asia, as well as the world’s largest giant observation wheel (Singapore Flyer) and oceanarium (Marine Life Park). Renowned as a world-class shopping paradise, Singapore also boasts an unrivalled selection of shopping options, the most famous being the Orchard Road cluster where the stores are open till 9.30pm on most nights and 11pm on Saturdays. For a unique evening shopping experience, scout around the ethnic quarters such as Chinatown and Little India, or even Mustafa Centre, which operates 24 hours a day, 365 days a year.

Come Taste the World

Due to her geographic advantage and rich multicultural heritage, the Lion City has become a true melting pot of flavours. You can see a reflection of Singapore’s cultural diversity in the array of local cuisines on the menu, creating its own uniquely Singapore flavor.

Chinese cuisine represents one of the main players in the country’s gastronomic arena. A visit to Singapore offers you an opportunity to sample dishes from the different parts of China. On your trip here, don’t forget to try local Chinese favourites such as chilli crab, bak kut teh, fish head curry or rojak. If you’re a fan of Indian food, you’ll be spoilt for choice between dishes from the southern and northern part of the sub-continent. You can also get a taste of popular local Indian-Muslim dishes that has been influenced along the years by the local Malay spices and cooking techniques. The Malay cuisine in Singapore will give you a chance to savour an array of spices and herbs including ginger, turmeric, galangal, lemon grass, curry leaves, pungent belachan (shrimp paste) and chillies. You’ll find the cuisine spicy without being unbearably hot, thanks to its generous use of coconut milk and local spices. The unique Peranakan or Nonya food offers a blend of Chinese, Malay and Indonesian flavours, combining aromatic herbs and spices such as lemongrass, chillies, tamarind paste, shrimp paste and coconut milk to create a rich cuisine of braised dishes, stews and curries.

And that’s far from all. Singapore also offers you a wide range of international cuisines – from Thai, Korean, Vietnamese to Mongolian food. Whether you’re in the mood for a Japanese dinner, a hearty Italian meal, or a casual French bistro experience, you’ll find it all in this little red dot.
INTERSPEECH 2014 General Information

Conference Venue
INTERSPEECH 2014 will be held in the MAX Atria @ Singapore EXPO. As part of Singapore EXPO, MAX Atria presents a brand-new conference experience. Designed around green initiatives and fitted with modern audio-visual technical capabilities, the venue provides dynamic spaces for any variety of events. MAX Atria is also the first exhibition and conference centre in Singapore to receive the BCA Green Mark Platinum award, signifying a commitment to sustainability efforts. The floor plan of MAX Atria can be found at the inside back cover.

Conference and Exhibition Hours

Registration:
Staff and student volunteers at the registration booths will be able to provide on-site assistance and advice.
FOYER, Level 1, MAX Atria, Singapore EXPO. Please see the floor plan for the exact location.
Saturday 13 September 2014  15:00 - 18:00
Sunday 14 September 2014  08:30 - 18:00
Monday 15 September 2014  07:30 - 19:00
Tuesday 16 September 2014  08:00 - 18:30
Wednesday 17 September 2014  08:00 - 18:30
Thursday 18 September 2014  08:00 - 15:30

Secretariat Office:
Staff will be available to assist and advise on conference-related and exhibition-related queries.
CITRINE (210 - 211), Level 2, MAX Atria. Please see the floor plan for the exact location.
Saturday 13 September 2014  15:00 - 18:00
Sunday - Wednesday 14 - 17 September 2014  08:00 - 18:30
Thursday 18 September 2014  08:00 - 17:00

Preview Area/Speaker Preparation:
All Speakers with a PowerPoint or PDF presentation must go to the Preview Area at least 60 minutes prior to their sessions. A team of AV and IT specialists will be ensuring that all presentations on USB sticks are in the correct format for use in the rooms and will be on hand to assist all Chairs and Speakers with any IT-related queries they have regarding their presentations.
CITRINE (210 - 211), Level 2, MAX Atria. Please see the floor plan for the exact location.
Saturday 13 September 2014  15:00 - 18:00
Sunday - Wednesday 14 - 17 September 2014  08:00 - 18:30
Thursday 18 September 2014  08:00 - 17:00

Name Badge
Please wear your conference badge at all times. Access will not be granted without it.

INTERSPEECH 2014 Conference App
This years’ edition of INTERSPEECH sees the extension of the conference application for smartphones and tablets. The app developed by the Quality and Usability Lab at TU Berlin, Germany, provides users with an up-to-date technical program, access to individual abstracts and papers, a personalized schedule which can be integrated into the smartphone calendar, search functionality including sessions, papers and authors, as well as selected general information about the conference venue and
local tourism. The app is available free of charge and supports Android (from version 4.0) and iOS (from Version 6.0). Full paper access is restricted to conference attendees.

Under the auspices of ISCA, Sebastian Möller and his team are developing the app into a platform for use in future INTERSPEECH and other ISCA events. Ultimately they hope to extend the functionality to test speech technology, i.e. voice search or synthesized speech output. You are invited to provide the team with user feedback and proposals for the development into a community platform, either at the INTERSPEECH 2015 booth, or via e-mail to interspeechapp@tu-berlin.de. Further information including links to download the app for your device, can be found at http://interspeechapp.qu.tu-berlin.de

Lost & Found
If you have lost or found anything, please contact the Secretariat Office.

Free WiFi
Powered by Xirrus, each array can support up to 200 concurrent users at a bandwidth of 2mbps securely, and you can stay connected where you are in Max Atria. For access code or connectivity-related query, staff at the registration booths or secretariat office will be able to assist and advice.

Disabled Access
There are lifts conveniently located near to the exhibition and conference rooms. Please see the floor plan for the exact location of these lifts.

Hospitality and Retail
Level 1, Singapore EXPO
Automated Teller Machines (ATMs), water stations, nursing room, restaurants and cafes are conveniently located near the exhibition and conference rooms. The Concierge team of Singapore EXPO located at Foyer, Level 1, MAX Atria, will be able to provide assistance and advice.

Business Centre
Level 1, Singapore EXPO
Monday - Friday: 09:00 - 18:00
Saturday: 09:00 - 13:00
Jointly managed by SingEx Venues and Starhub Ltd, the services available at the EXPO Business Centre include IT and telecommunication solutions, printing, faxing, photocopying, rental of telecommunication device/equipment, etc. Baggage storage is also available for hand carried bags and suitcases. For more information, please contact the staff at the Business Centre, e-mail: expobizctr@starhub.com, Hotline: 1800-822 3333, Fax: 6822 6002

Food Allergies
Labels are provided for the food served at the conference. Please contact the staff at the buffet tables for advice regarding the food and any particular requirement you may have.

First Aid
For first aid or emergency assistance, please contact any staff at MAX Atria for trained first aiders and cardiac first-responders on site. The Concierge team of Singapore EXPO located at Foyer, Level 1, MAX Atria, will be able to provide assistance and advice.
Taxi, Limousine Services
There are taxi stands conveniently located in front of MAX Atria, Hall 1 and next to Hall 6 of Singapore EXPO. For booking of the taxi or limousines services, please check with the Concierge staff at Foyer, Level 1, MAX Atria.

Currency
The local currency is the Singapore dollar (SGD). Money changing services can be found at the nearby shopping centres and hotels. ATMs accept most credit cards such as Visa, Mastercard and American Express.

Smoking Policies
The conference venue is a non-smoking area. Smoking is only possible in the designated yellow box outside MAX Atria, near Level 1 Opal rooms.
Social Program

Welcome Reception
*Monday, 15 September 2014, 19:00 - 21:00; Gallery, Level 1, MAX Atria (near exhibition area)*

Begin your conference experience with an evening of networking and fun while enjoying the food, drinks and entertainment. All attendees are welcome to join in on the celebration of the diversity of spoken languages and kick off the INTERSPEECH 2014 socially with a magical experience!

Please remember to bring your name badge and complimentary welcome reception ticket to access the welcome reception.

Student Reception
*Tuesday, 16 September 2014, 19:00 - 21:00; Topaz, Level 2, MAX Atria*

This student event is reserved to students. Entry tickets are provided with the student package and will be requested to access the student reception. The event is sponsored by Nuance Communications, Inc.

Students-Meet-Experts
*Wednesday, 17 September 2014, 12:00 - 13:30; Topaz, Level 2, MAX Atria*

Organized by ISCA SAC, partially sponsored by Google and Nuance Communications, Inc., students attending this event will be able to meet and discuss with experts and other students at the roundtables to bring home many interesting thoughts. It will definitely be an invaluable and enriching experience for the students.

Students must register with ISCA SAC to attend this event. Please wear your name badge and bring the ticket to participate in the students meet experts’ roundtables.

Banquet
*Wednesday, 17 September 2014, 19:30 - 22:30; Stamford Ballroom, Level 4, Raffles City Convention Centre (RCCC)*

Let your senses be treated to an evening of sumptuous dining and entertainment as we invite you to our INTERSPEECH 2014 Banquet in RCCC. Opening with an address from our General Conference Chair, Professor Haizhou Li, the evening will be followed by an exquisite Oriental 8-course dinner spread along with an entertaining line-up of ethnic drumming and singing. We look forward to spending an eventful evening with each and every one of our esteemed guests; we ensure that you will be well-fed and thoroughly engaged through the night!

Please remember to bring your name badge and complimentary banquet ticket to access the banquet.
Organizers

Organizing Committee

General Conference Chair
Haizhou LI (Institute for Infocomm Research, A*STAR, Singapore)

General Conference Co-Chair
Pak-Chung CHING (The Chinese University of Hong Kong, Hong Kong SAR, China)

Technical Program Chair
Helen MENG (The Chinese University of Hong Kong, Hong Kong SAR, China)

Technical Program Co-Chair
Bin MA (Institute for Infocomm Research, A*STAR, Singapore)

Local Arrangement Chair
Swee Lan SEE (Institute for Infocomm Research, A*STAR, Singapore)

Workshop Chair
Chai WUTIWIWATCHAI (NECTEC, Thailand)

Tutorial Chair
Eliathamby AMBIKAIRAJAH (University of New South Wales, Australia)

Plenary Session Chair
Brian MAK (Hong Kong University of Science and Technology, Hong Kong SAR, China)

Special Session Chair
Tomi KINNUNEN (University of Eastern Finland, Finland)

Publication Chair
Eng Siong CHNG (Nanyang Technological University, Singapore)
Lei XIE (Northwestern Polytechnical University, China)

Exhibition Chair
Xiong XIAO (Nanyang Technological University, Singapore)

Technology Show Chair
Alvina GOH (DSO National Laboratories, Singapore)

Publicity Chair
Anthony LARCHER (Institute for Infocomm Research, A*STAR, Singapore)

Sponsorship Chair
Nancy CHEN (Institute for Infocomm Research, A*STAR, Singapore)

Webmaster
Lei WANG (Institute for Infocomm Research, A*STAR, Singapore)

Local Arrangement Committee
Swee Lan SEE (Institute for Infocomm Research, A*STAR, Singapore)
Rong TONG (Institute for Infocomm Research, A*STAR, Singapore)
Kong Aik LEE (Institute for Infocomm Research, A*STAR, Singapore)
Siu Wa Yvonne LEE (Institute for Infocomm Research, A*STAR, Singapore)
Xi WANG (Institute for Infocomm Research, A*STAR, Singapore)
Dongyan HUANG (Institute for Infocomm Research, A*STAR, Singapore)
Cheung Chi LEUNG (Institute for Infocomm Research, A*STAR, Singapore)
Sunil SIVADAS (Institute for Infocomm Research, A*STAR, Singapore)
Yaozhu Paul CHAN (Institute for Infocomm Research, A*STAR, Singapore)
Boon Pang LIM (Institute for Infocomm Research, A*STAR, Singapore)
Hanwu SUN (Institute for Infocomm Research, A*STAR, Singapore)
Mei Hwan LOKE (Institute for Infocomm Research, A*STAR, Singapore)
Luis Fernando D’Haro ENRIQUEZ (Institute for Infocomm Research, A*STAR, Singapore)
Rafael E. BANCHS (Institute for Infocomm Research, A*STAR, Singapore)
Jochen Walter EHNES (Institute for Infocomm Research, A*STAR, Singapore)
Terence Wen Zheng NG (Institute for Infocomm Research, A*STAR, Singapore)
Su Jun LEOW (Nanyang Technological University, Singapore)
Guangpu HUANG (Nanyang Technological University, Singapore)
Nicholas DE LACZKOVICH-SIDDIQI (University of Applied Sciences, Berlin, Germany)

Advisory Committee

Jean-François BONASTRE (University of Avignon, France)
Pak-Chung CHING (The Chinese University of Hong Kong, Hong Kong SAR, China)
Hynek HERMANSKY (Johns Hopkins University, USA)
Chin-Hui LEE (Georgia Institute of Technology, USA)
Lin-shan LEE (National Taiwan University, Taiwan)
Haizhou LI (Institute for Infocomm Research, A*STAR, Singapore)
Satoshi NAKAMURA (Nara Institute of Science and Technology, Japan)
Douglas O'SHAUGHNESSY (INRS-EMT, University of Quebec, Canada)
Tanja SCHULTZ (Karlsruhe Institute of Technology, Germany)
Geok Leng TAN (A*STAR, Singapore)
Isabel TRANCOSO (INESC-ID, Portugal)
Chiu-Yu TSENG (Institute of Linguistics, Academia Sinica, Taiwan)
Technical Program Committee

Technical Program Chair
Helen MENG (The Chinese University of Hong Kong, Hong Kong SAR, China)

Technical Program Co-Chair
Bin MA (Institute for Infocom Research, A*STAR, Singapore)

Area Chairs

Area 1: Speech Perception and Production
Abeer ALWAN (University of California, Los Angeles, USA)
Jianwu DANG (Tianjin University, China)
Tan LEE (The Chinese University of Hong Kong, Hong Kong SAR, China)
Helmer STRIK (Radboud Universiteit Nijmegen, The Netherlands)

Area 2: Prosody, Phonetics, Phonology, and Para-/Non- Linguistic Information
Keikichi HIROSE (University of Tokyo, Japan)
Julia HIRSCHBERG (Columbia University, USA)
Bernd MÖBIUS (Saarland University, Germany)

Area 3: Analysis of Speech and Audio Signals
Hynek HERMANSKY (Johns Hopkins University, USA)
Malcolm SLANEY (Microsoft, USA)
Huy Dat TRAN (Institute for Infocomm Research, A*STAR, Singapore)

Area 4: Speech Coding and Enhancement
Rainer MARTIN (Ruhr-Universität Bochum, Germany)
Elmar NÖTH (University of Erlangen-Nuremberg, Germany)
Rongshan YU (Institute for Infocomm Research, A*STAR, Singapore)

Area 5: Speaker and Language Identification
William CAMPBELL (MIT Lincoln Laboratory, USA)
Kong Aik LEE (Institute for Infocomm Research, A*STAR, Singapore)
Man-Wai MAK (Hong Kong Polytechnic University, Hong Kong SAR, China)

Area 6: Speech Synthesis and Spoken Language Generation
Minghui DONG (Institute for Infocomm Research, A*STAR, Singapore)
Beat PFISTER (ETH Zurich, Switzerland)
Tomoki TODA (Nara Institute of Science and Technology, Japan)

Area 7: Speech Recognition - Signal Processing, Acoustic Modeling, Robustness, and Adaptation
Thomas HAIN (University of Sheffield, UK)
Mark HASEGAWA-JOHNSON (University of Illinois, USA)
Kate KNILL (University of Cambridge, UK)
Brian MAK (Hong Kong University of Science & Technology, Hong Kong SAR, China)

Area 8: Speech Recognition - Architecture, Search & Linguistic Components
Michiel BACCHIANI (Google, USA)
Thomas Fang ZHENG (Tsinghua University, China)
Area 9: LVCSR and Its Applications, Technologies and Systems for New Applications
  Eng Siong CHNG (Nanyang Technological University, Singapore)
  Lori LAMEL (LIMSI, France)
  Douglas O’SHAUGHNESSY (Université du Québec, Canada)
Area 10: Spoken Language Processing - Dialogue, Summarization, Understanding
  Joakim GUSTAFSON (KTH Royal Institute of Technology, Sweden)
  Giuseppe RICCARDI (University of Trento, Italy)
  Kai YU (Shanghai Jiaotong University, China)
Area 11: Spoken Language Processing – Translation, Info Retrieval
  Pascale FUNG (Hong Kong University of Science & Technology, Hong Kong SAR, China)
  Tatsuya KAWAHARA (Kyoto University Sakyo-ku, Japan)
  Hermann NEY (RWTH Aachen University, Germany)
Area 12: Spoken Language Evaluation, Standardization and Resources
  Isabel TRANCOSO (INESC-ID, Portugal)
  Jiahong YUAN (Linguistic Data Consortium, University of Pennsylvania, USA)

Special Session Chair
  Tomi KINNUNEN (University of Eastern Finland, Finland)

Plenary Session Chair
  Brian MAK (Hong Kong University of Science and Technology, Hong Kong SAR, China)

Tutorial Chair
  Eliathamby AMBIKAIRAJAH (University of New South Wales, Australia)

Workshop Chair
  Chai WUTIWIWATCHAI (NECTEC, Thailand)

Technology Show Chair
  Alvina GOH (DSO National Laboratories, Singapore)
<table>
<thead>
<tr>
<th>Scientific Review Committee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alberto Abad</td>
</tr>
<tr>
<td>Nassima Abdelli-Beruh</td>
</tr>
<tr>
<td>Andre Adami</td>
</tr>
<tr>
<td>Gilles Adda</td>
</tr>
<tr>
<td>Martine Adda-Decker</td>
</tr>
<tr>
<td>Yannis Agiomyrgiannakis</td>
</tr>
<tr>
<td>Masato Akagi</td>
</tr>
<tr>
<td>Masami Akamine</td>
</tr>
<tr>
<td>Paavo Alku</td>
</tr>
<tr>
<td>Alexander Allauzen</td>
</tr>
<tr>
<td>Fil Alleva</td>
</tr>
<tr>
<td>Jesús B. Alonso</td>
</tr>
<tr>
<td>Tanel Alumäe</td>
</tr>
<tr>
<td>Eliathamby Ambikairajah</td>
</tr>
<tr>
<td>Angélique Amelot</td>
</tr>
<tr>
<td>Kanae Amino</td>
</tr>
<tr>
<td>Noam Amir</td>
</tr>
<tr>
<td>Bistra Andreeva</td>
</tr>
<tr>
<td>Walter Andrews</td>
</tr>
<tr>
<td>Jorn Anemuller</td>
</tr>
<tr>
<td>Pontgip Angkitirakul</td>
</tr>
<tr>
<td>Xavier Anguera</td>
</tr>
<tr>
<td>Mark Antoniou</td>
</tr>
<tr>
<td>Takayuki Arai</td>
</tr>
<tr>
<td>Masahiro Araki</td>
</tr>
<tr>
<td>Yasuo Arikì</td>
</tr>
<tr>
<td>Ebru Arisoy</td>
</tr>
<tr>
<td>Aladdin Ariyaeenia</td>
</tr>
<tr>
<td>Denis Arnold</td>
</tr>
<tr>
<td>Hagai Aronowitz</td>
</tr>
<tr>
<td>Raman Arora</td>
</tr>
<tr>
<td>Victoria Arranz</td>
</tr>
<tr>
<td>Levent Arslan</td>
</tr>
<tr>
<td>Ramón Astudillo</td>
</tr>
<tr>
<td>Vincent Aubanel</td>
</tr>
<tr>
<td>Roland Auckenthaler</td>
</tr>
<tr>
<td>Nicolas Audibert</td>
</tr>
<tr>
<td>Cinzia Avesani</td>
</tr>
<tr>
<td>Tom Bäckström</td>
</tr>
<tr>
<td>Pierre Badin</td>
</tr>
<tr>
<td>Paolo Baggia</td>
</tr>
<tr>
<td>Odile Bagou</td>
</tr>
<tr>
<td>Gerard Bailly</td>
</tr>
<tr>
<td>Srinivas Bangalore</td>
</tr>
<tr>
<td>Plinio Barbosa</td>
</tr>
<tr>
<td>Ellen German Bard</td>
</tr>
<tr>
<td>Claude Barras</td>
</tr>
<tr>
<td>Vincent Barriac</td>
</tr>
<tr>
<td>Fernando Batista</td>
</tr>
<tr>
<td>Anton Batliner</td>
</tr>
<tr>
<td>Frederic Bechet</td>
</tr>
<tr>
<td>Steve Beet</td>
</tr>
<tr>
<td>Homayoon Beigi</td>
</tr>
<tr>
<td>Peter Bell</td>
</tr>
<tr>
<td>Patrice Bellot</td>
</tr>
<tr>
<td>Atef Ben Youssef</td>
</tr>
<tr>
<td>Jose Miguel Benedi</td>
</tr>
<tr>
<td>Stefan Benus</td>
</tr>
<tr>
<td>Mohamed Faouzi Benzeghibi</td>
</tr>
<tr>
<td>Nicole Beringer</td>
</tr>
<tr>
<td>Jared Bernstein</td>
</tr>
<tr>
<td>Frederic Berthommier</td>
</tr>
<tr>
<td>Laurent Besacier</td>
</tr>
<tr>
<td>Jonas Beskow</td>
</tr>
<tr>
<td>Michael Betser</td>
</tr>
<tr>
<td>Corine Bickley</td>
</tr>
<tr>
<td>Peter Birkholz</td>
</tr>
<tr>
<td>Jason Bishop</td>
</tr>
<tr>
<td>Judith Bishop</td>
</tr>
<tr>
<td>Alan W Black</td>
</tr>
<tr>
<td>Olivier Boëffard</td>
</tr>
<tr>
<td>Antonio Bonafonte</td>
</tr>
<tr>
<td>Jean-François Bonastre</td>
</tr>
<tr>
<td>Zinny Bond</td>
</tr>
<tr>
<td>Nandini Bondale</td>
</tr>
<tr>
<td>Anne Bonneau</td>
</tr>
<tr>
<td>Hynek Boril</td>
</tr>
<tr>
<td>Philippe Boula de Mareüil</td>
</tr>
<tr>
<td>Véronique Boullenger</td>
</tr>
<tr>
<td>Gilles Boulianne</td>
</tr>
<tr>
<td>Herve Bourlard</td>
</tr>
<tr>
<td>Suzanne Boyce</td>
</tr>
<tr>
<td>Daniela Braga</td>
</tr>
<tr>
<td>Bettina Braun</td>
</tr>
<tr>
<td>Hervé Bredin</td>
</tr>
<tr>
<td>Andrew Bren</td>
</tr>
<tr>
<td>John Bridle</td>
</tr>
<tr>
<td>Niko Brummer</td>
</tr>
<tr>
<td>Muratza Bulut</td>
</tr>
<tr>
<td>H Timothy Bunnell</td>
</tr>
<tr>
<td>Harry Bunt</td>
</tr>
<tr>
<td>Susanne Burger</td>
</tr>
<tr>
<td>Felix Burkhardt</td>
</tr>
<tr>
<td>Ian S. Burnett</td>
</tr>
<tr>
<td>Denis Burnham</td>
</tr>
<tr>
<td>Carlos Busso</td>
</tr>
<tr>
<td>Luis Caldas de Oliveira</td>
</tr>
<tr>
<td>Jose Ramon Calvo de Lara</td>
</tr>
<tr>
<td>Nick Campbell</td>
</tr>
<tr>
<td>Valentin Cardenoso-Payo</td>
</tr>
<tr>
<td>Patrick Cardinal</td>
</tr>
<tr>
<td>Michael Carey</td>
</tr>
<tr>
<td>Christopher Carignan</td>
</tr>
<tr>
<td>Rolf Carlson</td>
</tr>
<tr>
<td>Francisco Casacuberta</td>
</tr>
<tr>
<td>Diamantino Caseiro</td>
</tr>
<tr>
<td>Steve Cassidy</td>
</tr>
<tr>
<td>Maria Jose Castro-Bleda</td>
</tr>
<tr>
<td>Lawrence Cavedon</td>
</tr>
<tr>
<td>Asli Celikyilmaz</td>
</tr>
<tr>
<td>Christophe Cerisara</td>
</tr>
<tr>
<td>Delphine Charlet</td>
</tr>
<tr>
<td>Upendra Chaudhari</td>
</tr>
<tr>
<td>Ciprian Chelba</td>
</tr>
<tr>
<td>Berlin Chen</td>
</tr>
<tr>
<td>Gang Chen</td>
</tr>
<tr>
<td>Langzhou Chen</td>
</tr>
<tr>
<td>Nancy Chen</td>
</tr>
<tr>
<td>Xin Chen</td>
</tr>
<tr>
<td>Yan-Ming Cheng</td>
</tr>
<tr>
<td>Rathinavelu Chengalvarayan</td>
</tr>
<tr>
<td>Mohamed Chetouani</td>
</tr>
<tr>
<td>Jen-Tzung Chien</td>
</tr>
<tr>
<td>K.K. Chin</td>
</tr>
<tr>
<td>Ioana Chitoran</td>
</tr>
<tr>
<td>Gérard Chollet</td>
</tr>
<tr>
<td>Heidi Christensen</td>
</tr>
<tr>
<td>Ken Church</td>
</tr>
<tr>
<td>Robert Clark</td>
</tr>
<tr>
<td>Luisa Coheur</td>
</tr>
<tr>
<td>Alistair Conkic</td>
</tr>
<tr>
<td>Martin Cooke</td>
</tr>
<tr>
<td>Robin Cooper</td>
</tr>
<tr>
<td>Ricardo Cordoba</td>
</tr>
<tr>
<td>Martin Corley</td>
</tr>
<tr>
<td>Piero Cosi</td>
</tr>
<tr>
<td>Stephen Cox</td>
</tr>
<tr>
<td>Alex Cristia</td>
</tr>
<tr>
<td>Olivier Crouzet</td>
</tr>
<tr>
<td>Jia Cui</td>
</tr>
<tr>
<td>Xiaodong Cui</td>
</tr>
<tr>
<td>Fred Cummins</td>
</tr>
<tr>
<td>Francesco Cutugno</td>
</tr>
<tr>
<td>Deborah Dahl</td>
</tr>
<tr>
<td>Lirong Dai</td>
</tr>
<tr>
<td>Falavigna Daniele</td>
</tr>
<tr>
<td>Morena Danieli</td>
</tr>
<tr>
<td>Khalid Daoudi</td>
</tr>
<tr>
<td>Marelie Davel</td>
</tr>
<tr>
<td>Chris Davis</td>
</tr>
<tr>
<td>Amaro de Lima</td>
</tr>
<tr>
<td>Jose Mario De Martino</td>
</tr>
<tr>
<td>Renato de Mori</td>
</tr>
<tr>
<td>David Dean</td>
</tr>
<tr>
<td>Gilles Degottex</td>
</tr>
<tr>
<td>Najim Dehak</td>
</tr>
<tr>
<td>Michael Deisher</td>
</tr>
<tr>
<td>Veronique Delvaux</td>
</tr>
<tr>
<td>Grazyna Demenko</td>
</tr>
<tr>
<td>Kris Demuynek</td>
</tr>
<tr>
<td>Yasuhraru Den</td>
</tr>
<tr>
<td>Huiqun Deng</td>
</tr>
<tr>
<td>Li Deng</td>
</tr>
<tr>
<td>Jonathan Dennis</td>
</tr>
<tr>
<td>Anoop Deoras</td>
</tr>
<tr>
<td>Laurence Devillers</td>
</tr>
<tr>
<td>Luis Fernando D'Haro</td>
</tr>
<tr>
<td>Giuseppe Di Fabbrizio</td>
</tr>
<tr>
<td>Dimitrios Dimitriadis</td>
</tr>
<tr>
<td>Mariapaola D'Impero</td>
</tr>
<tr>
<td>Pierre Divenyi</td>
</tr>
<tr>
<td>Paul Dixon</td>
</tr>
<tr>
<td>Gerry Docherty</td>
</tr>
<tr>
<td>Laura Docio-Fernandez</td>
</tr>
<tr>
<td>Rama Sanand Doddpatla</td>
</tr>
<tr>
<td>Pierre Dognin</td>
</tr>
<tr>
<td>Hans Dolfing</td>
</tr>
<tr>
<td>Name</td>
</tr>
<tr>
<td>-----------------------------</td>
</tr>
<tr>
<td>Christoph Draxler</td>
</tr>
<tr>
<td>Joris Driesen</td>
</tr>
<tr>
<td>Carlo Drioli</td>
</tr>
<tr>
<td>Jasha Droppo</td>
</tr>
<tr>
<td>Thomas Drugman</td>
</tr>
<tr>
<td>Andrzej Drygajlo</td>
</tr>
<tr>
<td>Jacques Duchateau</td>
</tr>
<tr>
<td>Richard Dufour</td>
</tr>
<tr>
<td>Sophie Dufour</td>
</tr>
<tr>
<td>Stéphane Dupont</td>
</tr>
<tr>
<td>Thierry Dutoit</td>
</tr>
<tr>
<td>Jens Edlund</td>
</tr>
<tr>
<td>Robert Eklund</td>
</tr>
<tr>
<td>Daniel Ellis</td>
</tr>
<tr>
<td>Julien Epps</td>
</tr>
<tr>
<td>Hakan Erdogan</td>
</tr>
<tr>
<td>Donna Erickson</td>
</tr>
<tr>
<td>Daniel Erro</td>
</tr>
<tr>
<td>Engin Erzin</td>
</tr>
<tr>
<td>David Escudero</td>
</tr>
<tr>
<td>Maxine Ekenazi</td>
</tr>
<tr>
<td>Carol Espy-Wilson</td>
</tr>
<tr>
<td>Yannick Estève</td>
</tr>
<tr>
<td>Keelan Evaniini</td>
</tr>
<tr>
<td>Nicholas Evans</td>
</tr>
<tr>
<td>Florian Eynen</td>
</tr>
<tr>
<td>Mauro Falcone</td>
</tr>
<tr>
<td>Isabel Falé</td>
</tr>
<tr>
<td>Tiago Falk</td>
</tr>
<tr>
<td>Xing Fan</td>
</tr>
<tr>
<td>Qiang Fang</td>
</tr>
<tr>
<td>Jérôme Farinas</td>
</tr>
<tr>
<td>Kevin Farrell</td>
</tr>
<tr>
<td>David Farris</td>
</tr>
<tr>
<td>Marcos Faundez-Zanuy</td>
</tr>
<tr>
<td>Benoit Fauve</td>
</tr>
<tr>
<td>Benoit Favre</td>
</tr>
<tr>
<td>Tibor Fegyó</td>
</tr>
<tr>
<td>Sidney Fels</td>
</tr>
<tr>
<td>Junlan Feng</td>
</tr>
<tr>
<td>Raul Fernandez</td>
</tr>
<tr>
<td>Isabelle Ferrane</td>
</tr>
<tr>
<td>Marc Ferras</td>
</tr>
<tr>
<td>Javier Ferreiros</td>
</tr>
<tr>
<td>Carlos Ferrer</td>
</tr>
<tr>
<td>Luciana Ferrer</td>
</tr>
<tr>
<td>Tim Fingscheidt</td>
</tr>
<tr>
<td>Janet Fletcher</td>
</tr>
<tr>
<td>José A. R. Fonollosa</td>
</tr>
<tr>
<td>Eric Fosler-Lussier</td>
</tr>
<tr>
<td>Charles Fox</td>
</tr>
<tr>
<td>Robert A. Fox</td>
</tr>
<tr>
<td>Horacio Franco</td>
</tr>
<tr>
<td>Corinne Fredouille</td>
</tr>
<tr>
<td>Gerald Friedland</td>
</tr>
<tr>
<td>Guillaume Fuchs</td>
</tr>
<tr>
<td>Susanne Fuchs</td>
</tr>
<tr>
<td>Mark Fuhs</td>
</tr>
<tr>
<td>Masakiyo Fujimoto</td>
</tr>
<tr>
<td>Toshiaki Fukada</td>
</tr>
<tr>
<td>Takashi Fukuda</td>
</tr>
<tr>
<td>Sadaoki Furui</td>
</tr>
<tr>
<td>Name</td>
</tr>
<tr>
<td>-----------------------</td>
</tr>
<tr>
<td>Stefanie Jannedy</td>
</tr>
<tr>
<td>Aren Jansen</td>
</tr>
<tr>
<td>Peter Jax</td>
</tr>
<tr>
<td>Jesper Jensen</td>
</tr>
<tr>
<td>Alexandra Jesse</td>
</tr>
<tr>
<td>Luis M.T. Jesus</td>
</tr>
<tr>
<td>Hui Jiang</td>
</tr>
<tr>
<td>Qin Jin</td>
</tr>
<tr>
<td>Cheolwoo Jo</td>
</tr>
<tr>
<td>Michael Johnson</td>
</tr>
<tr>
<td>Kristina Jokinen</td>
</tr>
<tr>
<td>Caroline Jones</td>
</tr>
<tr>
<td>Arne Jonsson</td>
</tr>
<tr>
<td>Szu-Chen Stan Jou</td>
</tr>
<tr>
<td>Denis Jouvet</td>
</tr>
<tr>
<td>Preethi Jyothi</td>
</tr>
<tr>
<td>Tokihiko Kaburagi</td>
</tr>
<tr>
<td>Alexander Kain</td>
</tr>
<tr>
<td>Kaustubh Kalgaonkar</td>
</tr>
<tr>
<td>John Kane</td>
</tr>
<tr>
<td>Dimitri Kanevsky</td>
</tr>
<tr>
<td>Hong-Goo Kang</td>
</tr>
<tr>
<td>Stephan Kanthak</td>
</tr>
<tr>
<td>Martin Karafiat</td>
</tr>
<tr>
<td>Alexey Karpov</td>
</tr>
<tr>
<td>Hiroaki Kato</td>
</tr>
<tr>
<td>Athanasios Katsamanis</td>
</tr>
<tr>
<td>Hideki Kawahara</td>
</tr>
<tr>
<td>Shin'ichi Kawamoto</td>
</tr>
<tr>
<td>Patricia Keating</td>
</tr>
<tr>
<td>Patrick Kenny</td>
</tr>
<tr>
<td>Sameer ud Dowla Khan</td>
</tr>
<tr>
<td>Elie Khoury</td>
</tr>
<tr>
<td>Genichiro Kikui</td>
</tr>
<tr>
<td>Hoi Rin Kim</td>
</tr>
<tr>
<td>Hong Kook Kim</td>
</tr>
<tr>
<td>Hyung Soon Kim</td>
</tr>
<tr>
<td>Jeessun Kim</td>
</tr>
<tr>
<td>Samuel Kim</td>
</tr>
<tr>
<td>Sunhee Kim</td>
</tr>
<tr>
<td>Wool Kim</td>
</tr>
<tr>
<td>Owen Kimball</td>
</tr>
<tr>
<td>Brian Kingsbury</td>
</tr>
<tr>
<td>Norhide Kitaoka</td>
</tr>
<tr>
<td>Esther Klabbers</td>
</tr>
<tr>
<td>Dietrich Klakow</td>
</tr>
<tr>
<td>Anssi Klapuri</td>
</tr>
<tr>
<td>Felicitas Kleber</td>
</tr>
<tr>
<td>Bastiaan Kleijn</td>
</tr>
<tr>
<td>Takao Kobayashi</td>
</tr>
<tr>
<td>Tetsunori Kobayashi</td>
</tr>
<tr>
<td>Jachym Kolar</td>
</tr>
<tr>
<td>Kazunori Komatani</td>
</tr>
<tr>
<td>Mariko Kondo</td>
</tr>
<tr>
<td>Myoung-Wan Koo</td>
</tr>
<tr>
<td>Sunil Kumar Kopparapu</td>
</tr>
<tr>
<td>Takafulmi Koshinaka</td>
</tr>
<tr>
<td>Nir Krause</td>
</tr>
<tr>
<td>Nitish Krishnamurthy</td>
</tr>
<tr>
<td>Bernd Kroeger</td>
</tr>
<tr>
<td>Christian Kroeger</td>
</tr>
<tr>
<td>Roland Kuhn</td>
</tr>
<tr>
<td>Jimmy Kunzmann</td>
</tr>
<tr>
<td>Grace Kuo</td>
</tr>
<tr>
<td>Gakuto Kurata</td>
</tr>
<tr>
<td>Mikko Kurimo</td>
</tr>
<tr>
<td>Chul Hong Kwon</td>
</tr>
<tr>
<td>Oh-Wook Kwon</td>
</tr>
<tr>
<td>Francisco Lacerda</td>
</tr>
<tr>
<td>Pietro Laface</td>
</tr>
<tr>
<td>Catherine Lai</td>
</tr>
<tr>
<td>Ian Lane</td>
</tr>
<tr>
<td>Brian Langner</td>
</tr>
<tr>
<td>Yves Laprie</td>
</tr>
<tr>
<td>Anthony Larcher</td>
</tr>
<tr>
<td>Romain Laroche</td>
</tr>
<tr>
<td>Martha Larson</td>
</tr>
<tr>
<td>Kornel Laskowski</td>
</tr>
<tr>
<td>Javier Latorre</td>
</tr>
<tr>
<td>Jonathan Le Roux</td>
</tr>
<tr>
<td>Gwéolé Lecorvé</td>
</tr>
<tr>
<td>Chin-Hui Lee</td>
</tr>
<tr>
<td>Gary Geunbae Lee</td>
</tr>
<tr>
<td>Jaewon Lee</td>
</tr>
<tr>
<td>Lin-shan Lee</td>
</tr>
<tr>
<td>Siu-Wa Lee</td>
</tr>
<tr>
<td>Adrian Leemann</td>
</tr>
<tr>
<td>Fabrice Lefevre</td>
</tr>
<tr>
<td>Milan Legát</td>
</tr>
<tr>
<td>Yun Lei</td>
</tr>
<tr>
<td>Yi Ren Leng</td>
</tr>
<tr>
<td>Cheung-Chi Leung</td>
</tr>
<tr>
<td>Gary Leung</td>
</tr>
<tr>
<td>Michael Levit</td>
</tr>
<tr>
<td>Aijun Li</td>
</tr>
<tr>
<td>Feipeng Li</td>
</tr>
<tr>
<td>Jinyu Li</td>
</tr>
<tr>
<td>Junfeng Li</td>
</tr>
<tr>
<td>Kai Li</td>
</tr>
<tr>
<td>Ming Li</td>
</tr>
<tr>
<td>Hank Liao</td>
</tr>
<tr>
<td>Robin Lickley</td>
</tr>
<tr>
<td>Carlos Lima</td>
</tr>
<tr>
<td>Georges Linares</td>
</tr>
<tr>
<td>Anders Lindström</td>
</tr>
<tr>
<td>Zhen-Hua Ling</td>
</tr>
<tr>
<td>Diane Litman</td>
</tr>
<tr>
<td>Chaojun Liu</td>
</tr>
<tr>
<td>Gang Liu</td>
</tr>
<tr>
<td>Jian Liu</td>
</tr>
<tr>
<td>Jingjing Liu</td>
</tr>
<tr>
<td>Wenju Liu</td>
</tr>
<tr>
<td>Xunying (Andrew) Liu</td>
</tr>
<tr>
<td>Yang Liu</td>
</tr>
<tr>
<td>Eduardo Lleida Solano</td>
</tr>
<tr>
<td>Joaquim Lissterri</td>
</tr>
<tr>
<td>Anders Lofqivist</td>
</tr>
<tr>
<td>Damien Lolive</td>
</tr>
<tr>
<td>Yanhua Long</td>
</tr>
<tr>
<td>José Lopes</td>
</tr>
<tr>
<td>Ramon Lopez-Cozar</td>
</tr>
<tr>
<td>Teresa Lopez-Soto</td>
</tr>
<tr>
<td>Jorge Lucero</td>
</tr>
<tr>
<td>Steven Lulich</td>
</tr>
<tr>
<td>Susann Luperfoy</td>
</tr>
<tr>
<td>Jordi Luque</td>
</tr>
<tr>
<td>Changxue Ma</td>
</tr>
<tr>
<td>Chengyuan Ma</td>
</tr>
<tr>
<td>Ning Ma</td>
</tr>
<tr>
<td>Ewen MacDonald</td>
</tr>
<tr>
<td>Javier Macias-Guarasa</td>
</tr>
<tr>
<td>Ian Maddieson</td>
</tr>
<tr>
<td>Kikuo Maekawa</td>
</tr>
<tr>
<td>Hari Krishna Maganti</td>
</tr>
<tr>
<td>Mathew Magimai Doss</td>
</tr>
<tr>
<td>Ranniery Maia</td>
</tr>
<tr>
<td>Andreas Maier</td>
</tr>
<tr>
<td>Nicolas Malyska</td>
</tr>
<tr>
<td>Nuno Mamede</td>
</tr>
<tr>
<td>Michael Mandel</td>
</tr>
<tr>
<td>Robert Mannell</td>
</tr>
<tr>
<td>Kazunori Mano</td>
</tr>
<tr>
<td>Krzysztof Marasek</td>
</tr>
<tr>
<td>Stefania Marin</td>
</tr>
<tr>
<td>David Marks</td>
</tr>
<tr>
<td>Jean-Pierre Martens</td>
</tr>
<tr>
<td>David Martinez Gonzalez</td>
</tr>
<tr>
<td>Carlos-D. Martinez-Hinarejos</td>
</tr>
<tr>
<td>David Martins de Matos</td>
</tr>
<tr>
<td>John Mason</td>
</tr>
<tr>
<td>Takashi Masuko</td>
</tr>
<tr>
<td>Ana Isabel Mata</td>
</tr>
<tr>
<td>Marco Matassoni</td>
</tr>
<tr>
<td>Pavel Matejka</td>
</tr>
<tr>
<td>Tomoko Matsu</td>
</tr>
<tr>
<td>Yuri Matveev</td>
</tr>
<tr>
<td>Ludo Max</td>
</tr>
<tr>
<td>Catherine Mayo</td>
</tr>
<tr>
<td>Alan McCree</td>
</tr>
<tr>
<td>Erik McDermott</td>
</tr>
<tr>
<td>Mitchell McLaren</td>
</tr>
<tr>
<td>Michael McTear</td>
</tr>
<tr>
<td>Daryush Mehta</td>
</tr>
<tr>
<td>Sylvain Meignier</td>
</tr>
<tr>
<td>Hugo Meinedo</td>
</tr>
<tr>
<td>Aleksandro Meireles</td>
</tr>
<tr>
<td>Einar Meister</td>
</tr>
<tr>
<td>Angeliki Metallinou</td>
</tr>
<tr>
<td>Florian Metze</td>
</tr>
<tr>
<td>Christine Meunier</td>
</tr>
<tr>
<td>Yohann Meynadier</td>
</tr>
<tr>
<td>Ben Milner</td>
</tr>
<tr>
<td>Yasuhiro Minami</td>
</tr>
<tr>
<td>Nobuaki Minematsu</td>
</tr>
<tr>
<td>Taniya Mishra</td>
</tr>
<tr>
<td>Ananya Misra</td>
</tr>
<tr>
<td>Vikramjit Mitra</td>
</tr>
<tr>
<td>Hansjörg Mixdorff</td>
</tr>
<tr>
<td>Takemi Mochida</td>
</tr>
<tr>
<td>Parham Mokhtari</td>
</tr>
<tr>
<td>Helena Moniz</td>
</tr>
<tr>
<td>Juan M Montero</td>
</tr>
<tr>
<td>Elliot Moore</td>
</tr>
<tr>
<td>Roger Moore</td>
</tr>
<tr>
<td>Sylvia Moosmueller</td>
</tr>
<tr>
<td>Nicolas Morales</td>
</tr>
</tbody>
</table>
Abhinav Sethy
Izhak Shafran
Stefanie Shattuck-Hufnagel
Jason Shaw
Slava Shechtman
Koichi Shinoda
Takahiro Shinozaki
Ryan Shosted
Elizabeth Shirberg
Jan Silovska
Carlos Silva
Elliot Singer
Olivier Siohan
Man-hung Siu
Sunil Sivadas
Gabriel Skantze
Jan Skoglund
Raymond Slyh
Kamel Smaili
Paris Smaragdis
Connie So
Rudolph Sock
Hagen Soltau
Frank Soong
Victor Sorokin
Ian Stavness
Stefan Steidl
Ingmar Steiner
Evgeny Stepanov
Andreas Stolcke
Brad Story
Svetlana Stoyanchev
Sofia Strömbergsson
Brian Strope
Jane Stuart-Smith
Sebastian Stüker
Douglas Sturim
David Suendermann
Jun-Won Suh
Rafid Sukkar
Xie Sun
Masayuki Suzuki
Piergiorgio Svaizer
Torbjörn Svendsen
Marc Swerts
Ann Syrdal
Igor Szoke
Marija Tabain
Yuuki Tachioka
Satoshi Takahashi
Toru Takahashi
Tetsuya Takiguchi
Yik-Cheung Tam
Fabio Tamburini
Zheng-Hua Tan
Kazuyo Tanaka
Yun Tang
Jianhua Tao
Arvi Tavast
António Teixeira
João Paulo Teixeira
Louis ten Bosch
Fabio Tesser
Barry-John Theobald
Samuel Thomas
Michael Tjalve
Roberto Togneri
Shinichi Tokuma
Markus Toman
Laura Tomokiyō
Rong Tong
Pedro Torres-Carrasquillo
Asterios Toutios
David Traum
Juegen Trouvain
Khiest Truong
Yu Tsao
Chiu-yu Tseng
Sh-Chuan Tseng
Kimiko Tsukada
Gokhan Tur
Oyrun Turk
Markku Turunen
Michael Tyler
Masashi Unoki
Michel Vacher
Martti Vainio
Claudio Vair
Dirk Van Compernolle
Rogier van Dalen
Henk van den Heuvel
Laurens van der Werff
Hugo Van hamme
David van Leeuwen
Maarten Van Segbroeck
Rob van Son
Amparo Varona
Ioana Vasilescu
Eric Vatikiotis-Bateson
Nanette Veilleux
Werner Verhelst
Klara Vici
Marina Vigario
Coriandré Vilain
Jesús Villalba
Fernando Villavicencio
Emmanuel Vincent
Oriol Vinyals
Ravichander Vipperla
Tuomas Virtanen
Carl Vogel
Stephen Voran
Anil Kumar Vuppala
Petra Wagner
Michael Walsh
Patrick Wambacq
Gsiao-Chuan Wang
Hsin-Min Wang
Xinhao Wang
Yongqiang Wang
Nigel Ward
Wayne Ward
Shinji Watanabe
Catherine Watson
Jianguo Wei
Eugene Weinstein
Pauline Welby
Christian Wellekens
Chao Weng
Stanley Wenndt
Stefan Werner
Mirjam Wester
Daniel Willett
Jason D Williams
Mats Wirén
Guillaume Wisniewski
Chuck Wooters
Monika Woszczyna
Johan Wouters
Chung-Hsien Wu
Zhiyong Wu
Zhizheng Wu
Chai Wutiwimatchai
Xiong Xiao
Jian Xue
Junichi Yamagishi
Yoichi Yamashita
Yonghong Yan
Zhi-Jie Yan
Bayya Yegnanarayana
Hani Camille Yehia
Sree Harsha Yella
Serdar Yildirim
Nestor Becerra Yoma
Kiyoko Yoneyama
Chang Yoo
Dongsuk Yook
Changhuai You
Steve Young
Dong Yu
Meng Yu
Young-Sun Yun
Francois Yvon
Stephen Zahorian
Ilija Zeljkovic
Heiga Zen
Andrej Zgank
Chi (Leo) Zhang
Pengyuan Zhang
Yunxin Zhao
Xinhui Zhou
Wolffram Ziegler
Frank Zimmerer
Imed Zitouni
Udo Zoelzer
Enrico Zovato
Geoffrey Zweig
Marzena Zygis
Future INTERSPEECH Conferences

SEPTMBER 6 - 10
DRESDEN, GERMANY

Conference
Speech is the most important biosignal humans can produce and perceive. Still, not all characteristics of speech are fully understood. INTERSPEECH 2015 will therefore be organized around the topic “Speech: Towards a Better Understanding of the Most Important Biosignal.” The event will be staged in the recently built Martin International Congress Center (ICC) in Dresden, the capital of Saxony. It is located in the heart of Europe, easily reached via Berlin or Dresden airports, and will offer a great deal of history and culture to INTERSPEECH 2015 delegates.

INTERSPEECH 2015 emphasizes an interdisciplinary approach covering all aspects of speech science and technology spanning basic theories to applications. In addition to regular oral and poster sessions, the conference will also feature plenary talks by internationally renowned experts, tutorials, special sessions, show & tell sessions, and exhibits. A number of satellite events will take place immediately before and after the conference. Please follow the details which will be published at interspeech2015.org

Calls
31.08.2014
Proposals for satellite workshops

Upcoming Calls
15.10.2014 - Call for Papers
15.10.2014 - Call for Tutorials
15.10.2014 - Call for Special Sessions

Location
Martins International Congress Center
Dresden (ICD)
Gosta-Ufer 2
01067 Dresden
Germany
Tel: +49 351 216 0
Fax: +49 351 216 1000
Web: dresden-congresscenter.de

Further questions?
http://interspeech2015.org

Understanding Speech Processing in Humans and Machines

SAVE THE DATE!
September 8-12, 2016

The Hyatt Regency
San Francisco, California

General Chair:
Nelson Morgan

Technical Chairs:
Shrikanth Narayanan
Panayiotis Georgiou
Satellite Events
The INTERSPEECH 2014 workshop committee chaired by Dr. Chai Wutiwiwatchai is pleased to announce the following six satellite events, which are scheduled either before or after the main conference.

**MA3HMI 2014: 2nd Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction**
14 September 2014
Location: Singapore
Website https://www.scss.tcd.ie/conferences/MA3HMI/index.html

**Blizzard Challenge 2014: Blizzard Challenge Workshop 2014**
19 September 2014
Location: Singapore
Website http://synsig.org/index.php/Blizzard_Challenge_2014_Workshop

**ISCSLP 2014: 9th International Symposium on Chinese Spoken Language Processing**
12-14 September 2014
Location: Singapore
Website http://www.iscslp2014.org/

**WOCCI 2014: 4th Workshop on Child, Computer and Interaction**
19 September 2014
Location: Singapore
Website http://www.wocci.org/

**O-COCOSDA 2014: 17th Oriental COCOSDA**
10-12 September 2014
Location: Phuket, Thailand
Website http://www.ococosda2014.org/

**SLAM 2014: 2nd Workshop on Speech, Language and Audio in Multimedia**
11-12 September 2014
Location: Penang, Malaysia
Website http://language.cs.usm.my/SLAM2014/
Keynote Speeches

Anne CUTLER

ISCA Medalist 2014

MARCS Institute, University of Western Sydney, Australia

Monday, 15 September 2014, 09:30 - 10:30; Garnet 213 - 218, Level 2, MAX Atria

Keynote Speech 1: Learning About Speech

Abstract

We human language users learn about speech all our lives. In fact if the beginning of our individual life is taken to be the moment of our birth, we learn for more than our whole lives, because even prior to birth we acquire much basic information about language sound structure. Then in early life, before we are capable of uttering any recognizable words, we have already learned a huge amount about the sounds and words of our language; furthermore, the efficiency with which we learn this is predictive of our later linguistic facility. We learn to process speech in the way best suited to the particular language (or languages) we acquire as children. This leads to a formidable efficiency and robustness in native-language processing, but it actually inhibits learning in the case of other languages we encounter only later - children are clearly very much better at learning a new language than adults are! Learning about speech nonetheless continues throughout life, in particular the everyday perceptual learning that enables us to adapt our speech processing to newly encountered talkers, and to adjust our own pronunciation in keeping with pronunciation changes in our speech community across time. This learning (at least in the native language) can draw on a wide variety of information sources, is fully in place in childhood, and is apparently unattenuated in older language users.

Biography

Anne Cutler is professor in the MARCS Institute, University of Western Sydney, and Processing program leader of the newly established ARC Centre of Excellence in the Dynamics of Language. She studied in Australia, Germany and the US, and worked in the UK (Sussex, Cambridge) and from 1993 to 2013 as director at the Max Planck Institute for Psycholinguistics in Nijmegen, The Netherlands. Her research, of which her book Native Listening (MIT Press 2012) gives an overview, centres on human listeners’ recognition of spoken language. It has tended over the years to involve a great many cross-linguistic comparisons (e.g., English, Dutch, German, Japanese, Cantonese, Korean, Sesotho, French, Spanish, Italian, Finnish, Polish, Arabic, Telugu, Berber - so far).
Keynote Speech 2: Decision Learning in Data Science: Where John Nash Meets Social Media

Abstract
With the increasing ubiquity and power of mobile devices, as well as the prevalence of social media, more and more activities in our daily life are being recorded, tracked, and shared, creating the notion of “social media”. Such abundant and still growing real life data, known as “big data”, provide a tremendous research opportunity in many fields. To analyze, learn and understand such user-generated big data, machine learning has been an important tool and various machine learning algorithms have been developed. However, since the user-generated big data is the outcome of users’ decisions, actions and their socio-economic interactions, which are highly dynamic, without considering users’ local behaviors and interests, existing learning approaches tend to focus on optimizing a global objective function at the macroeconomic level, while totally ignore users’ local decisions at the microeconomic level. As such there is a growing need in bridging machine/social learning with strategic decision making, which are two traditionally distinct research disciplines, to be able to jointly consider both global phenomenon and local effects to understand/model/analyze better the newly arising issues in the emerging social media. In this talk, we present the notion of “decision learning” that can involve users’ behaviors and interactions by combining learning with strategic decision making. We will discuss some examples from social media with real data to show how decision learning can be used to better analyze users’ optimal decision from a user’s perspective as well as design a mechanism from the system designer’s perspective to achieve a desirable outcome.

Biography
Dr. K. J. Ray Liu was named a Distinguished Scholar-Teacher of University of Maryland in 2007, where he is Christine Kim Eminent Professor of Information Technology. He leads the Maryland Signals and Information Group conducting research encompassing broad areas of signal processing and communications with recent focus on cooperative communications, cognitive networking, social learning and decision making, and information forensics and security. Dr. Liu has received numerous honors and awards including IEEE Signal Processing Society 2009 Technical Achievement Award and various best paper awards from IEEE Signal Processing, Communications, and Vehicular Technology Societies, and EURASIP. A Fellow of the IEEE and AAAS, he is recognized by Thomson Reuters as an ISI Highly Cited Researcher. Dr. Liu was the President of IEEE Signal Processing Society, the Editor-in-Chief of IEEE Signal Processing Magazine and the founding Editor-in-Chief of EURASIP Journal on Advances in Signal Processing. Dr. Liu also received various research and teaching recognitions from the University of Maryland, including Poole and Kent Senior Faculty Teaching Award, Outstanding Faculty Research Award, and Outstanding Faculty Service Award, all from A. James Clark School of Engineering; and Invention of the Year Award (three times) from Office of Technology Commercialization.
Keynote Speech 3: Language Diversity: Speech Processing in A Multi-Lingual Context

Abstract
Speech processing encompasses a variety of technologies that automatically process speech for some downstream processing. These technologies include identifying the language or dialect spoken, the person speaking, what is said and how it is said. The downstream processing may be limited to a transcription or to a transcription enhanced with additional metadata, or may be used to carry out an action or interpreted within a spoken dialog system or more generally for analytics. With the availability of large spoken multimedia or multimodal data there is growing interest in using such technologies to provide structure and random access to particular segments. Automatic tools can also serve to annotate large corpora for exploitation in linguistic studies of spoken language, such as acoustic-phonetics, pronunciation variation and diachronic evolution, permitting the validation of hypotheses and models. In this talk I will present some of my experience with speech processing in multiple languages, drawing upon progress in the context of several research projects, most recently the Quaero program and the IARPA Babel program, both of which address the development of technologies in a variety of languages, with the aim to some highlight recent research directions and challenges.

Biography
I am a senior research scientist (DR1) at the CNRS, which I joined as a permanent researcher at LIMSI in October 1991. I received my Ph.D. degree in Electrical Engineering and Computer Science in May 1988 from the Massachusetts Institute of Technology. My research activities focus on large vocabulary speaker-independent, continuous speech recognition in multiple languages with a recent focus on low-resourced languages; lightly and unsupervised acoustic model training methods; studies in acoustic-phonetics; lexical and pronunciation modeling. I contributed to the design, and realization of large speech corpora (TIMIT, BREF, TED). I have been actively involved in the research projects, most recently leading the activities on speech processing in the OSEO Quaero program, and I am currently co-principal investigator for LIMSI as part of the IARPA Babel Babelon team led by BBN. I served on the Steering committee for INTERSPEECH 2013 as co-technical program chair along with Pascal Perrier, and I am now serving on the Technical Program Committee of INTERSPEECH 2014.
Abstract
In contrast to other species, humans are unique in having developed thousands of diverse languages which are not mutually intelligible. However, any infant can learn any language with ease, because all languages are based upon common biological infrastructures of sensori-motor, memorial, and cognitive faculties. While languages may differ significantly in the sounds they use, the overall organization is largely the same. It is divided into a discrete segmental system for building words and a continuous prosodic system for expressing, phrasing, attitudes, and emotions. Within this organization, I will discuss a class of languages called 'tone languages', which makes special use of F0 to build words. Although the best known of these is Chinese, tone languages are found in many parts of the world, and operate on different principles. I will also comment on relations between sound patterns in language and sound patterns in music, the two worlds of sound universal to our species.

Speaker’s Bio
Dr. William S-Y. Wang received his early schooling in China, and his Ph.D. from the University of Michigan. He was appointed Professor of Linguistics at the University of California at Berkeley in 1965, and taught there for 30 years. Currently he is in the Department of Electronic Engineering and in the Department of Linguistics and Modern Languages of the Chinese University of Hong Kong, and Director of the newly established Joint Research Centre for Language and Human Complexity. His primary interest is the evolution of language from a multi-disciplinary perspective.
Abstract
Artificial neural networks have been around for over half a century and their applications to speech processing have been almost as long, yet it was not until year 2010 that their real impact had been made by a deep form of such networks, built upon part of the earlier work on (shallow) neural nets and (deep) graphical models developed by both speech and machine learning communities. This keynote will first reflect on the path to this transformative success, sparked by speech analysis using deep learning methods on spectrogram-like raw features and then progressing rapidly to speech recognition with increasingly larger vocabularies and scale. The role of well-timed academic-industrial collaboration will be highlighted, so will be the advances of big data, big compute, and the seamless integration between the application-domain knowledge of speech and general principles of deep learning. Then, an overview will be given on sweeping achievements of deep learning in speech recognition since its initial success in 2010 (as well as in image recognition and computer vision since 2012). Such achievements have resulted in across-the-board, industry-wide deployment of deep learning. The final part of the talk will look ahead towards stimulating new challenges of deep learning - making intelligent machines capable of not only hearing (speech) and seeing (vision), but also of thinking with a “mind”; i.e. reasoning and inference over complex, hierarchical relationships and knowledge sources that comprise a vast number of entities and semantic concepts in the real world based in part on multi-sensory data from the user. To this end, language and multimodal processing - joint exploitation and learning from text, speech/audio, and image/video - is evolving into a new frontier of deep learning, beginning to be embraced by a mixture of research communities including speech and spoken language processing, natural language processing, computer vision, machine learning, information retrieval, cognitive science, artificial intelligence, and data/knowledge management. A review of recent published studies will be provided on deep learning applied to selected language and multimodal processing tasks, with a trace back to the relevant early connectionist modeling and neural network literature and with future directions in this new exciting deep learning frontier discussed and analyzed.

Speaker’s Bio
Dr. Li Deng received his Ph.D. from the University of Wisconsin-Madison. He was a tenured professor (1989 - 1999) at the University of Waterloo, Ontario, Canada, and then joined Microsoft Research, Redmond, where he is currently a Principal Research Manager of its Deep Learning Technology Center. Since 2000, he has also been an affiliate full professor at the University of Washington, Seattle, teaching computer speech processing. He has been granted over 60 US or international patents, and has received numerous awards and honors bestowed by IEEE, ISCA, ASA, and Microsoft including the latest IEEE SPS Best Paper Award (2013) on deep neural nets for speech recognition. He authored or co-authored 4 books including the latest one on Deep Learning: Methods and Applications. He is a Fellow of the Acoustical Society of America, a Fellow of the IEEE, and a Fellow of the ISCA. He served as the Editor-in-Chief for IEEE Signal Processing Magazine (2009 - 2011), and currently as Editor-in-Chief for IEEE Transactions on Audio, Speech and Language Processing. His recent research interests and activities have been focused on deep learning and machine intelligence applied to large-scale text analysis and to speech/language/image multimodal processing, advancing his earlier work with collaborators on speech analysis and recognition using deep neural networks since 2009.
Special Sessions

SP 1: Open Domain Situated Conversational Interaction

*Monday, 15 September 2014, 11:00 - 13:00; Peridot 206, Level 2, MAX Atria*

Robust conversational systems have the potential to revolutionize our interactions with computers. Building on decades of academic and industrial research, we now talk to our computers, phones, and entertainment systems on a daily basis. However, current technology typically limits conversational interactions to a few narrow domains/topics (e.g., weather, traffic, restaurants). Users increasingly want the ability to converse with their devices over broad web-scale content. Finding something on your PC or the web should be as simple as having a conversation. A promising approach to address this problem is situated conversational interaction. The approach leverages the situation and/or context of the conversation to improve system accuracy and effectiveness. Sources of context include visual content being displayed to the user, geo-location, prior interactions, multi-modal interactions (e.g., gesture, eye gaze), and the conversation itself. For example, while a user is reading a news article on their tablet PC, they initiate a conversation to dig deeper on a particular topic. Or a user is reading a map and wants to learn more about the history of events at mile marker 121. Or a gamer wants to interact with a game’s characters to find the next clue in their quest. All of these interactions are situated – rich context is available to the system as a source of priors/constraints on what the user is likely to say. This special session will provide a forum to discuss research progress in open domain situated conversational interactions. Topics of the session will include:

- Situated context in spoken dialog systems
- Visual/dialog/personal/geo situated context
- Inferred context through interpretation and reasoning
- Open domain spoken dialog systems
- Open domain spoken/natural language understanding and generation
- Open domain semantic interpretation
- Open domain dialog management (large-scale belief state/policy)
- Conversational interactions
- Multi-modal inputs in situated open domains (speech/text + gesture, touch, eye gaze)
- Multi-human situated interactions

Organizers:
Larry Heck (larry@ieee.org), Microsoft Research
Dilek Hakkani-Tür (dilek@ieee.org), Microsoft Research
Gokhan Tur (gokhan@ieee.org), Microsoft Research
Steve Young (sjy@eng.cam.ac.uk), Cambridge University

SP 2: INTERSPEECH 2014 Computational Paralinguistics Challenge (ComParE)

*Monday, 15 September 2014, 14:30 - 16:30; Peridot 206, Level 2, MAX Atria*

*Monday, 15 September 2014, 17:00 - 19:00; Peridot 206, Level 2, MAX Atria*

The INTERSPEECH 2014 Computational Paralinguistics Challenge (ComParE) is an open challenge dealing with speaker characteristics as manifested in their speech signal's acoustic properties. This year, it introduces new tasks by the Cognitive Load Sub-Challenge and the Physical Load Sub-Challenge. For these tasks, the COGNITIVE-LOAD WITH SPEECH AND EGG database (CLSE) and the MUNICH BIOVOICE CORPUS (MBC) with high diversity of speakers and different languages covered (Australian English and German) are provided by the organizers. The corpora contain fully realistic data in challenging acoustic conditions and feature rich annotation such as speaker meta-data. They are given with distinct definitions of test, development, and training partitions; speaker independence is guaranteed as needed in most real-life settings. Benchmark results of the most popular approaches are provided as in the years before. The transcription of the train and development sets is known. All Sub-Challenges allow contributors to find their own features with their own machine learning algorithm. However, a standard feature set has been provided per corpus that could be used. Participants had to stick to the definition of training, development, and test sets. They may report on results obtained on the development set, but had only five trials to upload their results on the test sets, whose labels are unknown to them. Participants had to submit a paper presenting the results.
that underwent peer-review and had to be accepted for the conference in order to participate in the Challenge. The results of the Challenge are presented in a Special Session (two time slots) at INTERSPEECH 2014 in Singapore.

Organizers:
Björn Schuller (schuller@IEEE.org), Imperial College London / Technische Universität München
Stefan Steidl (stefan.steidl@fau.de), Friedrich-Alexander-University
Anton Batliner (batliner@cs.fau.de), Technische Universität München / Friedrich-Alexander-University
Jarek Krajweski (krajewsk@uni-wuppertal.de), Bergische Universität Wuppertal
Julien Epps (j.epps@unsw.edu.au), The University of New South Wales / National ICT Australia

SP 3: Speech Technologies for Ambient Assisted Living
Tuesday, 16 September 2014, 10:00 - 12:00; Peridot 206, Level 2, MAX Atria

This special session focuses on the use of speech technologies for ambient assisted living, the creation of smart spaces and intelligent companions that can preserve independence and executive function, social communication and security of people with special needs. Currently, speech interfaces for assistive technologies remains underutilized despite its potential to replace or augment obtrusive and sometimes outright inaccessible conventional computer interfaces. Moreover in a smart home context, efficiency of speech interfaces can be supported by a number of concurrent information sources (e.g., wearable sensors, home automation sensors), enabling multimodal communication. In practice, daily hand-free usage of speech interfaces remains limited due to challenging real-world conditions, and because conventional speech interfaces can have difficulty with the atypical speech of many users. This, in turn, can be attributed to the lack of abundant speech material, and the limited adaptation to the user of these systems. Taking up the challenges of this domain requires a multidisciplinary approach to define the user's needs, record corpora in realistic usage conditions, develop speech interfaces that are robust to both environment and user's characteristics and are able to adapt to specific users. This special session will bring together researchers in speech and audio technologies with people from the ambient assisted living and assistive technologies communities to meet and foster awareness between members of either community, discuss problems, techniques and datasets, and perhaps initiate common projects.

Organizers:
Heidi Christensen (h.christensen@dcs.shef.ac.uk), University of Sheffield
Jort F. Gemmeke (jgemmeke@amadana.nl), KU Leuven
François Portet (francois.portet@imag.fr), Laboratoire d'Informatique de Grenoble
Frank Rudzicz (frank@cs.toronto.edu), University of Toronto
Michel Vacher (michel.vacher@imag.fr), Laboratoire d'Informatique de Grenoble

SP 4: Text-Dependent Speaker Verification with Short Utterances
Tuesday, 16 September 2014, 15:00 - 17:00; Peridot 206, Level 2, MAX Atria

In recent years, speaker verification engines have reached maturity and have been deployed in commercial applications. Ergonomics of such applications is especially demanding and imposes a drastic limitation in terms of speech duration during authentication. A well known tactic to address the problem of lack of data, due to short duration, is using text-dependency. However, recent breakthroughs achieved in the context of text-independent speaker verification in terms of accuracy and robustness do not benefit text-dependent applications. Indeed, large development data required by the recent approaches is not available in the text-dependent context. The purpose of this special session is to gather the research efforts from both academia and industry toward a common goal of establishing a new baseline and explore new directions for text-dependent speaker verification. The focus of the session is on robustness with respect to duration and modeling of lexical information. To support the development and evaluation of text-dependent speaker verification technologies, the Institute for Infocomm Research, A*STAR, Singapore, has recently released the RSR2015 database, including 150 hours of data recorded from 300 speakers. Further details are available at: http://www1.i2r.a-star.edu.sg/~kalee/is2014/tdspk.html
Organizers:
Anthony Larcher (alarcher@i2r.a-star.edu.sg), Institute for Infocomm Research, A*STAR, Singapore
Hagai Aronowitz (hagaia@il.ibm.com), IBM Research – Haifa
Kong Aik Lee (kalee@i2r.a-star.edu.sg), Institute for Infocomm Research, A*STAR, Singapore
Patrick Kenny (patrick.kenny@crim.ca), CRIM – Montréal

SP 5: Phase Importance in Speech Processing Applications
Wednesday, 17 September 2014, 10:00 - 12:00; Peridot 206, Level 2, MAX Atria

In the past decades, the amplitude of the speech spectrum has been considered to be the most important feature for speech processing applications and phase of the speech signal has received less attention. Recently, several findings justify the phase importance in speech and audio processing communities. The importance of phase estimation along with amplitude estimation in speech enhancement, complementary phase-based features in speech and speaker recognition and phase-aware acoustic modeling of the environment are the most prominent reported works scattered in different communities of speech and audio processing. These examples suggest that incorporating the phase information can push the limits of the state-of-the-art phase-independent solutions employed for long in different aspects of audio and speech signal processing. This special session aims to explore the recent advances and methodologies to exploit the knowledge of signal phase information in different aspects of speech processing. Without a dedicated effort to bring researchers from different communities, a quick advance in investigation towards the phase usefulness in speech processing applications is difficult to achieve. Therefore, as the first step in this direction, we aim to promote the “phase-aware speech and audio signal processing” to form a community of researchers to organize the next steps. Our initiative is to unify these efforts to better understand the pros and cons of using phase and the degree of feasibility for phase estimation/enhancement in different areas of speech processing including: speech enhancement, speech separation, speech quality estimation, speech and speaker recognition, voice transformation and speech analysis and synthesis. The goal is to promote the importance of the phase-based signal processing and studying its importance and sharing interesting findings from different speech processing applications.

Organizers:
Pejman Mowlaee (pejman.mowlaee@tugraz.at), Graz University of Technology
Rahim Saeidi (rahim.saeidi@uef.fi), University of Eastern Finland
Yannis Stylianou (yannis@csd.uoc.gr), Toshiba Labs Cambridge UK / University of Crete

SP 6: Deep Neural Networks for Speech Generation and Synthesis
Wednesday, 17 September 2014, 13:30 - 15:30; Peridot 206, Level 2, MAX Atria
Wednesday, 17 September 2014, 16:00 - 18:00; Peridot 206, Level 2, MAX Atria

This special session aims to bring together researchers who work actively on deep neural network (DNN) for speech research, particularly, in generation and synthesis, to promote and to understand the state-of-art DNN research in statistical learning and compare results with the parametric HMM-GMM model based TTS synthesis, generation, and conversion. DNN, with its neuron-like structure, can simulate human speech production system in a layered, hierarchical, nonlinear and self-organized network. It can transform linguistic text information into intermediate semantic, phonetic and prosodic content and finally generate speech waveforms. Many possible neural network architectures or typologies exist, e.g. feed-forward NN with multiple hidden layers, stacked RBM or CRBM, Recurrent Neural Net (RNN), which have been used to speech/image recognition and other applications. We would like to use this special session as a forum to present updated results in the research frontiers, algorithm development and application scenarios.

Organizers:
Yao Qian (yaoqian@microsoft.com), Microsoft Research
Frank K. Soong (frankkps@microsoft.com), Microsoft Research
Special Events

Show & Tell (Demo)

Session 1: Tuesday, 16 September 2014, 15:00 - 17:00, Garnet Foyer, Level 2, MAX Atria
Session 2: Wednesday, 17 September 2014, 13:30 - 15:30, Garnet Foyer, Level 2, MAX Atria

Show & Tell is a special event organized during the conference, where participants are given the opportunity to demonstrate their most recent progress, developments or innovation tracks, and interact with the conference attendees in an informal way, such as a talk, a poster, a mockup, a demo, or any adapted format of their own choice. These contributions usually (but do not have to) relate to a paper and they must highlight the innovative side of the concept.

Show & Tell participants have submitted a specific 2 page written contribution, together with optional multimedia content. Each proposition for the Show & Tell sessions has been peer-reviewed. Evaluation was based on the originality, significance, quality, and clarity of each submission.


Students-Meet-Experts (Students’ Roundtable)

Wednesday, 17 September 2014, 12:00 - 13:30, Topaz, Level 2, MAX Atria

Students-Meet-Experts (SME) is a lunch event organized for student participants of INTERSPEECH. This is an occasion where students meet face to face with well known researchers and scientists in academia and industry while enjoying a nice meal. Students and senior members are arranged in a mixed seating to facilitate the interaction.

SME has become part of INTERSPEECH conference in the past years. It is organized by ISCA Student Advisory Committee with advice from ISCA Board's Student Liaison, Professor Helen Meng. SME is sponsored by both ISCA Board and industry partners. In 2014, we would like to thank Google and Nuance for their generous support.
The INTERSPEECH 2014 tutorial committee chaired by Professor Eliathamby Ambikairajah is pleased to announce the following eight tutorials at the conference and will be offered on Sunday, 14 September 2014. All tutorials will be of three (3) hours duration.

<table>
<thead>
<tr>
<th>Morning Tutorials</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>T1</strong></td>
<td><strong>Non-Speech Acoustic Event Detection And Classification</strong></td>
</tr>
<tr>
<td></td>
<td><em>Tuomas Virtanen and Jort F. Gemmeke</em></td>
</tr>
<tr>
<td></td>
<td><em>Sunday, 14 September 2014, 09:30 - 12:30; Peridot 201, Level 2, MAX Atria</em></td>
</tr>
<tr>
<td><strong>T2</strong></td>
<td><strong>Contribution of MRI to Exploring and Modeling Speech Production</strong></td>
</tr>
<tr>
<td></td>
<td><em>Kiyoshi Honda and Jianwu Dang</em></td>
</tr>
<tr>
<td></td>
<td><em>Sunday, 14 September 2014, 09:30 - 12:30; Peridot 202, Level 2, MAX Atria</em></td>
</tr>
<tr>
<td><strong>T3</strong></td>
<td><strong>Computational Models for Audiovisual Emotion Perception</strong></td>
</tr>
<tr>
<td></td>
<td><em>Emily Mower Provost and Carlos Busso</em></td>
</tr>
<tr>
<td></td>
<td><em>Sunday, 14 September 2014, 09:30 - 12:30; Peridot 205, Level 2, MAX Atria</em></td>
</tr>
<tr>
<td><strong>T4</strong></td>
<td><strong>The Art and Science of Speech Feature Engineering</strong></td>
</tr>
<tr>
<td></td>
<td><em>Samuel Thomas and Sriram Ganapathy</em></td>
</tr>
<tr>
<td></td>
<td><em>Sunday, 14 September 2014, 09:30 - 12:30; Peridot 206, Level 2, MAX Atria</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Afternoon Tutorials</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>T5</strong></td>
<td><strong>Recent Advances in Speaker Diarization</strong></td>
</tr>
<tr>
<td></td>
<td><em>Hagai Aronowitz</em></td>
</tr>
<tr>
<td></td>
<td><em>Sunday, 14 September 2014, 14:00 - 17:00; Peridot 201, Level 2, MAX Atria</em></td>
</tr>
<tr>
<td><strong>T6</strong></td>
<td><strong>Multimodal Speech Recognition with the AusTalk 3D Audio-Visual Corpus</strong></td>
</tr>
<tr>
<td></td>
<td><em>Roberto Togneri, Mohammed Bennamoun and Chao Sui</em></td>
</tr>
<tr>
<td></td>
<td><em>Sunday, 14 September 2014, 14:00 - 17:00; Peridot 202, Level 2, MAX Atria</em></td>
</tr>
<tr>
<td><strong>T7</strong></td>
<td><strong>Semantic Web and Linked Big Data Resources for Spoken Language Processing</strong></td>
</tr>
<tr>
<td></td>
<td><em>Dilek Hakkani-Tür and Larry Heck</em></td>
</tr>
<tr>
<td></td>
<td><em>Sunday, 14 September 2014, 14:00 - 17:00; Peridot 205, Level 2, MAX Atria</em></td>
</tr>
<tr>
<td><strong>T8</strong></td>
<td><strong>Speech and Audio for Multimedia Semantics</strong></td>
</tr>
<tr>
<td></td>
<td><em>Florian Metze and Koichi Shinoda</em></td>
</tr>
<tr>
<td></td>
<td><em>Sunday, 14 September 2014, 14:00 - 17:00; Peridot 206, Level 2, MAX Atria</em></td>
</tr>
</tbody>
</table>
T1: Non-Speech Acoustic Event Detection and Classification
Tuomas Virtanen (Tampere University of Technology, Finland) and Jort F. Gemmeke (KU Leuven, Belgium)
Sunday, 14 September 2014, 09:30 - 12:30; Peridot 201, Level 2, MAX Atria

Abstract: The research in audio signal processing has been dominated by speech research, but most of the sounds in our real-life environments are actually non-speech events such as cars passing by, wind, warning beeps, and animal sounds. These acoustic events contain much information about the environment and physical events that take place in it, enabling novel application areas such as safety, health monitoring and investigation of biodiversity. But while recent years have seen widespread adoption of applications such as speech recognition and song recognition, generic computer audition is still in its infancy. Non-speech acoustic events have several fundamental differences to speech, but many of the core algorithms used by speech researchers can be leveraged for generic audio analysis. The tutorial is a comprehensive review of the field of acoustic event detection as it currently stands. The goal of the tutorial is foster interest in the community, highlight the challenges and opportunities and provide a starting point for new researchers. We will discuss what acoustic event detection entails, the commonalities differences with speech processing, such as the large variation in sounds and the possible overlap with other sounds. We will then discuss basic experimental and algorithm design, including descriptions of available databases and machine learning methods. We will then discuss more advanced topics such as methods to deal with temporally overlapping sounds and modeling the relations between sounds. We will finish with a discussion of avenues for future research.

T2: Contribution of MRI to Exploring and Modeling Speech Production
Kiyoshi Honda (Tianjin University, China) and Jianwu Dang (JAIST, Japan)
Sunday, 14 September 2014, 09:30 - 12:30; Peridot 202, Level 2, MAX Atria

Abstract: Magnetic Resonance Imaging (MRI) provides us a magic vision to look into the human body in various ways not only with static imaging but also with motion imaging. MRI has been a powerful technique for speech research to study finer anatomy of the speech organs or to visualize true vocal tracts in three dimensions. Inherent problems of slow image acquisition for speech tasks or insufficient signal-to-noise ratio for microscopic observation have been the cost for researchers to search for task-specific imaging techniques. The recent advances of the 3-Tesla technology suggest more practical solutions to broader applications of MRI by overcoming previous technical limitations. In this joint tutorial in two parts, we summarize our previous effort to accumulate scientific knowledge with MRI and to advance speech modeling studies for future development. Part 1, given by Kiyoshi Honda, introduces how to visualize the speech organs and vocal tracts by presenting techniques and data for finer static imaging, synchronized motion imaging, surface marker tracking, real-time imaging, and vocal-tract mechanical modeling. Part 2, presented by Jianwu Dang, focuses on applications of MRI for phonetics of Mandarin vowels, acoustics of the vocal tracts with side branches, analysis and simulation in search of talker characteristics, physiological modeling of the articulatory system, and motor control paradigm for speech articulation.
T3: Computational Models for Audiovisual Emotion Perception
Emily Mower Provost (University of Michigan, USA) and Carlos Busso (University of Texas, Dallas, USA)
Sunday, 14 September 2014, 09:30 - 12:30; Peridot 205, Level 2, MAX Atria

Abstract: In this tutorial we will explore engineering approaches to understanding human emotion perception, focusing both on modeling and application. We will highlight both current and historical trends in emotion perception modeling, focusing on both psychological and engineering-driven theories of perception (statistical analyses, data-driven computational modeling, and implicit sensing). The importance of this topic can be appreciated from both an engineering viewpoint, any system that either models human behavior or interacts with human partners must understand emotion perception as it fundamentally underlies and modulates our communication, or from a psychological perspective, emotion perception is also used in the diagnosis of many mental health conditions and is tracked in therapeutic interventions. Research in emotion perception seeks to identify models that describe the felt sense of ‘typical’ emotion expression – i.e., an observer/evaluator’s attribution of the emotional state of the speaker. This felt sense is a function of the methods through which individuals integrate the presented multimodal emotional information. We will cover psychological theories of emotion, engineering models of emotion, and experimental approaches to measure emotion. We will demonstrate how these modeling strategies can be used as a component of emotion classification frameworks and how they can be used to inform the design of emotional behaviors.

T4: The Art and Science of Speech Feature Engineering
Sriram Ganapathy and Samuel Thomas (IBM T.J. Watson Research Center, USA)
Sunday, 14 September 2014, 09:30 - 12:30; Peridot 206, Level 2, MAX Atria

Abstract: With significant advances in mobile technology and audio sensing devices, there is a fundamental need to describe vast amounts of audio data in terms of well representative lower dimensional descriptors for efficient automatic processing. The extraction of these signal representations, also called features, constitutes the first step in processing a speech signal. The art and science of feature engineering relates to addressing the two inherent challenges - extracting sufficient information from the speech signal for the task at hand and suppressing the unwanted redundancies for computational efficiency and robustness. The area of speech feature extraction combines a wide variety of disciplines like signal processing, machine learning, psychophysics, information theory, linguistics and physiology. It has a rich history spanning more than five decades and has seen tremendous advances in the last few years. This has propelled the transition of the speech technology from controlled environments to millions of end user applications. In this tutorial, we review the evolution of speech feature processing methods, summarize the recent advances of the last two decades and provide insights into the future of feature engineering. This will include the discussions on the spectral representation methods developed in the past, human auditory motivated techniques for robust speech processing, data driven unsupervised features like iVectors and recent advances in deep neural network based techniques. With experimental results, we will also illustrate the impact of these features for various state-of-the-art speech processing systems. The future of speech signal processing will need to address various robustness issues in complex acoustic environments while being able to derive useful information from big data.
**T5: Recent Advances in Speaker Diarization**  
_Hagai Aronowitz (IBM Research, Haifa, Israel)_  
**Sunday, 14 September 2014, 14:00 - 17:00; Peridot 201, Level 2, MAX Atria**

**Abstract:** The tutorial will start with an introduction to speaker diarization giving a general overview of the subject. Afterwards, we will cover the basic background including feature extraction, and common modeling techniques such as GMMs and HMMs. Then, we will discuss the first processing step usually done in speaker diarization which is voice activity detection. We will consequently describe the classic approaches for speaker diarization which are widely used today. We will then introduce state-of-the-art techniques in speaker recognition required to understand modern speaker diarization techniques. Following, we will describe approaches for speaker diarization using advanced representation methods (supervectors, speaker factors, iVectors) and we will describe supervised and unsupervised learning techniques used for speaker diarization. We will also discuss issues such as coping with unknown number of speakers, detecting and dealing with overlapping speech, diarization confidence estimation, and online speaker diarization. Finally we will discuss two recent works: exploiting a-priori acoustic information (such as processing a meeting when some of the participants are known in advanced to the system, and training data is available for them), The second recent work is modeling speaker-turn dynamics. If time permits, we will also discuss concepts such as multi-modal diarization and using TDOA (time difference of arrival) for diarization of meetings.

---

**T6: Multimodal Speech Recognition with the AusTalk 3D Audio-Visual Corpus**  
_Roberto Togneri, Mohammed Bennamoun, and Chao (Luke) Sui (University of Western Australia, Australia)_  
**Sunday, 14 September 2014, 14:00 - 17:00; Peridot 202, Level 2, MAX Atria**

**Abstract:** This tutorial will provide attendees a brief overview of 3D based AVSR research. In this tutorial, attendees will learn how to use the newly developed 3D based audio visual data corpus we derived from the AusTalk corpus (https://austalk.edu.au/) for audio-visual speech/speaker recognition. In addition, we also plan to introduce some results using this newly developed 3D audio-visual data corpus, which show that there is a significant speech accuracy increase by integrating both depth-level and grey-level visual features. In the first part of the tutorial, we will review some recent works published in the last decade, so that attendees can obtain an overview of the fundamental concepts and challenges in this field. In the second part of the tutorial, we will briefly describe the recording protocol and contents of the 3D data corpus, and show attendees how to use this corpus for their own research. In the third part of this tutorial, we will present our results using the 3D data corpus. The experimental results show that, compared with the conventional AVSR based on the audio and grey-level visual features, the integration of grey and depth visual information can boost the AVSR accuracy significantly. Moreover, we will also experimentally explain why adding depth information can benefit the standard AVSR systems. Eventually, through our tutorial, we hope we can inspire more researchers in the community to contribute to this exciting research.
T7: Semantic Web and Linked Big Data Resources for Spoken Language Processing
Dilek Hakkani-Tür and Larry Heck (Microsoft Research, USA)
Sunday, 14 September 2014, 14:00 - 17:00; Peridot 205, Level 2, MAX Atria

Abstract: State-of-the-art statistical spoken language processing typically requires significant manual effort to construct domain-specific schemas (ontologies) as well as manual effort to annotate training data against these schemas. At the same time, a recent surge of activity and progress on semantic web-related concepts from the large search-engine companies represents a potential alternative to the manually intensive design of spoken language processing systems. Standards such as schema.org have been established for schemas (ontologies) that webmasters can use to semantically and uniformly markup their web pages. Search engines like Bing, Google, and Yandex have adopted these standards and are leveraging them to create semantic search engines at the scale of the web. As a result, the open linked data resources and semantic graphs covering various domains (such as Freebase [3]) have grown massively every year and contain far more information than any single resource anywhere on the Web. Furthermore, these resources contain links to text data (such as Wikipedia pages) related to the knowledge in the graph. Recently, several studies on speech language processing started exploiting these massive linked data resources for language modeling and spoken language understanding. This tutorial will include a brief introduction to the semantic web and the linked data structure, available resources, and querying languages. An overview of related work on information extraction and language processing will be presented, where the main focus will be on methods for learning spoken language understanding models from these resources.

T8: Speech and Audio for Multimedia Semantics
Florian Metze (Carnagie Mellon University, USA) and Koichi Shinoda (Tokyo Institute of Technology, Japan)
Sunday, 14 September 2014, 14:00 - 17:00; Peridot 206, Level 2, MAX Atria

Abstract: Internet media sharing sites and the one-click upload capability of smartphones are producing a deluge of multimedia content. While visual features are often dominant in such material, acoustic and speech information in particular often complements it. By facilitating access to large amounts of data, the text-based Internet gave a huge boost to the field of natural language processing. The vast amount of consumer-produced video becoming available now will do the same for video processing, eventually enabling semantic understanding of multimedia material, with implications for human computer interaction, robotics, etc. Large-scale multi-modal analysis of audio-visual material is now central to a number of multi-site research projects around the world. While each of these has slightly different targets, they are facing largely the same challenges: how to robustly and efficiently process large amounts of data, how to represent and then fuse information across modalities, how to train classifiers and segmenters on unlabeled data, how to include human feedback, etc. In this tutorial, we will present the state of the art in large-scale video, speech, and non-speech audio processing, and show how these approaches are being applied to tasks such as content-based video retrieval (CBVR) and multimedia event detection (MED). We will introduce the most important tools and techniques, and show how the combination of information across modalities can be used to induce semantics on multimedia material through ranking of information and fusion. Finally, we will discuss opportunities for research that the INTERSPEECH community specifically will find interesting and fertile.
Awards

ISCA Medal
The ISCA Medal 2014 will be awarded to Professor Anne Cutler by the President of ISCA during the opening ceremony.

ISCA Best Paper Awards
The ISCA Best Paper Awards 2014 will be announced by the President of ISCA during the closing ceremony.

ISCA Best Student Paper Award
Each year, ISCA awards 3 best student papers at INTERSPEECH based on anonymous reviewing and presentation at the conference. The ISCA Best Student Paper Awards are sponsored by IBM Research, and Dilek Hakkani-Tür, Larry Heck from Microsoft Research. Dilek Hakkani-Tür and Larry Heck are donating their honorarium from their tutorial “Semantic Web and Linked Big Data Resources for Spoken Language Processing” to sponsor the ISCA Best Student Paper Award.

This year, 12 papers are shortlisted for best student paper:

138 Investigating the Effect of F0 and Vocal Intensity on Harmonic Magnitudes: Data from Laryngeal High-Speed Video Endoscopy
   Gang Chen, Soo Jin Park, Jody Kreiman and Abeer Alwan
   Poster Session 16: Phonetics and Phonology
   Wednesday, 17 September 2014, 10:00 – 12:00

162 Automatic Estimation of the Lip Radiation Effect in Glottal Inverse filtering
   Manu Airaksinen, Tom Bäckström and Paavo Alku
   Oral Session 8: Speech Production I
   Monday, 15 September 2014, 14:30 – 16:30

361 Modeling Therapist Empathy through Prosody in Drug Addiction Counseling
   Bo Xiao, Daniel Bone, Maarten Van Segbroeck, Zac Imel, David Atkins, Panayiotis Georgiou and Shrikanth Narayanan
   Poster Session 2: Extraction of Para-Linguistic Information
   Monday, 15 September 2014, 11:00 – 13:00

497 Intrinsic Spectral Analysis Based on Temporal Context Features for Query by Example Spoken Term Detection
   Peng Yang, Cheung-Chi Leung, Lei Xie, Bin Ma and Haizhou Li
   Poster Session 17: Spoken Term Detection and Document Retrieval
   Wednesday, 17 September 2014, 10:00 – 12:00

517 Direct F0 Control of an Electrolarynx Based on Statistical Excitation Feature Prediction and its Evaluation through Simulation
   Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura and Ko Tanaka
   Oral Session 2: Prosody Processing
   Monday, 15 September 2014, 11:00 – 13:00

557 Towards a Neural Measure of Perceptual Distance - Classification of Electroencephalographic Responses to Synthetic Vowels
   Manson Cheuk-Man Fong, James William Minett, Thierry Blu and William Shi-Yuan Wang
   Oral Session 37: Speech Perception
   Thursday, 18 September 2014, 10:00 – 12:00
Exploring Modulation Spectrum Features for Speech-Based Depression Level Classification
Elif Bozkurt, Orith Toledo - Ronen, Alexander Sorin and Ron Hoory
Oral Session 18: Paralinguistic and Extralinguistic Information
Tuesday, 16 September 2014, 15:00 – 17:00

Lexical Representation of Consonant, Vowels and Tones in Early Childhood
Hwee Hwee Goh, Charlene Fu and Kheng Hui Yeo
Oral Session 36: Language Acquisition
Thursday, 18 September 2014, 10:00 – 12:00

Speech synthesis in various communicative situations: Impact of pronunciation variations
Sandrine Brognaux, Benjamin Picart and Thomas Drugman
Oral Session 21: Statistical Parametric Speech Synthesis
Wednesday, 17 September 2014, 10:00 – 12:00

Adaptive Speech Recognition and Dialogue Management for Users with Speech Disorders
Iñigo Casanueva, Heidi Christensen, Thomas Hain and Phil Green
Special Session 3: Speech Technologies for Ambient Assisted Living
Tuesday, 16 September 2014, 10:00 – 12:00

Word-level Invariant Representations from Acoustic Waveforms
Stephen Voinea, Chiyuan Zhang, Georgios Evangelopoulos, Lorenzo Rosasco and Tomaso Poggio
Poster Session 23: Speech Representation, Detection & Classification
Wednesday, 17 September 2014, 16:00 - 18:00

Acoustic Modeling with Deep Neural Networks Using Raw Time Signal for LVCSR
Zoltán Tüske, Pavel Golik, Ralf Schlüter and Hermann Ney
Oral Session 13: Feature Extraction and Modelling for ASR
Tuesday, 16 September 2014, 10:00 – 12:00
INTERSPEECH Travel Awards
Five INTERSPEECH Travel Awards have been selected based on the technical quality of the papers. The INTERSPEECH 2014 Organizing Committee thanks the sponsors Google (3), iFLYTEK Co., Ltd. (1), and the corporate partner Samsung (1) for their generosity.

This year, the 5 recipients for the INTERSPEECH Travel Awards are:

**Cyrielle Chappuis (Switzerland)**

396  When Voices Get Emotional: A Study of Emotion-Enhanced Memory and Impairment During Emotional Prosody Exposure
Oral Session 28: Perception of Emotion and Prosody
Wednesday, 17 September 2014, 13:30 – 15:30

**Polina Drozdova (The Netherlands)**

883  Phoneme Category Retuning in a Non-native Language
Poster Session 5: Cross-linguistic Studies
Monday, 15 September 2014, 14:30 – 16:30

**Diandra Fabre (France)**

934  Automatic Animation of an Articulatory Tongue Model from Ultrasound Images Using Gaussian Mixture Regression
Poster Session 22: Speech Synthesis II
Wednesday, 17 September 2014, 16:00 – 18:00

**Guangting Mai (UK)**

351  Relative importance of AM and FM cues for speech comprehension: Effects of speaking rate and their implications for neurophysiological processing of speech
Oral Session 37: Speech Perception
Thursday, 18 September 2014, 10:00 – 12:00

**Victor Soto (USA)**

1143 A Comparison of Multiple Methods for Rescoring Keyword Search Lists for Low Resource Languages
Oral Session 33: Spoken Term Detection for Low-Resource Languages I
Thursday, 18 September 2014, 10:00 – 12:00

ISCA Best Paper Published in Speech Communication (2011 - 2013)
The best paper has been selected by the Speech Communication editorial board under the coordination of Bernd Möbius, Editor-in-Chief.

The ISCA best Speech Communication paper award will be presented at INTERSPEECH2014 and announced during EUSIPCO 2014.

ISCA Best Paper Published in Computer Speech and Language (2009 - 2013)
The best paper has been selected by the Computer Speech and Language editorial board under the coordination of Roger K. Moore, Editor-in-Chief.
Sunday, 14 September 2014

Daily Schedule

8:00 - 12:00: Peridot 201 - Peridot 206
- 8:00: Registration
- 9:00: Tutorial 1: Non-speech Acoustic Event Detection and Classification
- 10:00: Tutorial 2: Contribution of MRI to Exploring and Modeling Speech Production
- 11:00: Tutorial 3: Computational Models for Audiovisual Emotion Perception
- 12:00: Tutorial 4: The Art and Science of Speech Feature Engineering
- 13:00: Lunch Break
- 14:00: Tutorial 5: Recent Advances in Speaker Diarization
- 15:00: Tutorial 6: Multimodal Speech Recognition with the AusTalk 3D Audio-Visual Corpus
- 16:00: Tutorial 7: Semantic Web and Linked Big Data Resources for Spoken Language Processing
- 17:00: Registration
- 18:00: Coffee Break

Coffee Break is from 11:00AM - 11:30AM for the morning sessions, and from 15:30PM - 16:00PM for the afternoon sessions.
<table>
<thead>
<tr>
<th>Time</th>
<th>Location</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:00</td>
<td>Garnet 213-218</td>
<td>Keynote 2: K. J. Ray Liu</td>
</tr>
<tr>
<td>09:00</td>
<td>Garnet 213-218</td>
<td>Coffee Break</td>
</tr>
<tr>
<td>10:00</td>
<td>Peridot 202-203</td>
<td>Oral 13: Feature Extraction and Modeling for ASR</td>
</tr>
<tr>
<td></td>
<td>Peridot 204-205</td>
<td>Oral 15: Speech Technologies and Applications</td>
</tr>
<tr>
<td></td>
<td>Peridot 201</td>
<td>Oral 16: Source Separation and Computational Auditory Scene Analysis</td>
</tr>
<tr>
<td></td>
<td>Peridot 206</td>
<td>Special 3: Speech Technologies for Ambient Assisted Living</td>
</tr>
<tr>
<td></td>
<td>Max Atria Gallery</td>
<td>Poster Sessions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10. DNN for ASR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11. Speaker Recognition - General Topics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12. Speech Processing with Multi-modalities</td>
</tr>
<tr>
<td>11:00</td>
<td></td>
<td>Oral 14: Speech Analysis</td>
</tr>
<tr>
<td>12:00</td>
<td></td>
<td>Oral 16: Source Separation and Computational Auditory Scene Analysis</td>
</tr>
<tr>
<td>13:00</td>
<td></td>
<td>Special 3: Speech Technologies for Ambient Assisted Living</td>
</tr>
<tr>
<td>14:00</td>
<td>Garnet 213-218</td>
<td>Keynote 3: Lori Lamel</td>
</tr>
<tr>
<td>15:00</td>
<td></td>
<td>Coffee Break</td>
</tr>
<tr>
<td>16:00</td>
<td></td>
<td>Oral 17: Normalization and Discriminative Training Methods</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oral 18: Paralinguistic and Extralinguistic Information</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oral 19: Text Processing for Speech Synthesis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oral 20: Cross-language Perception and Production</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Special 4: Text-dependent Speaker Verification With Short Utterances</td>
</tr>
<tr>
<td></td>
<td>Garnet Foyer</td>
<td>Poster Sessions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13. Speech and Audio Analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15. Pronunciation Modeling and Learning</td>
</tr>
<tr>
<td>18:00</td>
<td>Garnet 213-218</td>
<td>ISCA General Assembly</td>
</tr>
<tr>
<td>19:00</td>
<td>Topaz</td>
<td>Student Reception</td>
</tr>
<tr>
<td>20:00</td>
<td></td>
<td>Student Reception</td>
</tr>
<tr>
<td>21:00</td>
<td></td>
<td>Student Reception</td>
</tr>
</tbody>
</table>
Session Index

Monday 15 September 2014

Keynote 1: Anne Cutler, ISCA Medalist ................................................................. 61
Mon-Oral-1 Oral Session 1: Multi-Lingual ASR ......................................................... 61
Mon-Oral-2 Oral Session 2: Prosody Processing ....................................................... 62
Mon-Oral-3 Oral Session 3: Speaker Recognition — Applications ......................... 63
Mon-Oral-4 Oral Session 4: Phonetics and Phonology .............................................. 64
Mon-SP-1 Special Session 1: Open Domain Situated Conversational Interaction ....... 65
Mon-Poster-1 Poster Session 1: Speech Production: Models and Acoustics .............. 66
Mon-Poster-2 Poster Session 2: Extraction of Para-Linguistic Information ................ 69
Mon-Poster-3 Poster Session 3: Spoken Language Understanding ............................ 70
Mon-Oral-5 Oral Session 5: Spoken Dialogue Systems ............................................. 73
Mon-Oral-6 Oral Session 6: DNN Architectures and Robust Recognition ................. 74
Mon-Oral-7 Oral Session 7: Speaker Recognition — Evaluation and Forensics .......... 75
Mon-Oral-8 Oral Session 8: Speech Production I .................................................... 76
Mon-SP-2(a) Special Session 2(a): INTERSPEECH 2014 Computational Paralinguistics ChallengE (ComParE) I ................................................................. 77
Mon-Poster-4 Poster Session 4: Hearing and Perception .......................................... 78
Mon-Poster-5 Poster Session 5: Cross-Linguistic Studies ......................................... 81
Mon-Poster-6 Poster Session 6: Speaker Diarization ................................................ 83
Mon-Oral-9 Oral Session 9: Robust ASR .............................................................. 85
Mon-Oral-10 Oral Session 10: Implementation of Language Model Algorithms ........... 86
Mon-Oral-11 Oral Session 11: Speaker Recognition — Noise and Channel Robustness . 87
Mon-Oral-12 Oral Session 12: Speech Production II ................................................. 89
Mon-SP-2(b) Special Session 2(b): INTERSPEECH 2014 Computational Paralinguistics ChallengE (ComParE) II ................................................................. 90
Mon-Poster-7 Poster Session 7: Speech Synthesis I .................................................. 91
Mon-Poster-8 Poster Session 8: Multi-Lingual Cross-Lingual and Low-Resource ASR . 93
Mon-Poster-9 Poster Session 9: Speech Estimation and Sound Source Separation ...... 95

Tuesday 16 September 2014

Keynote 2: K. J. Ray Liu .................................................................................. 97
Tue-Oral-13 Oral Session 13: Feature Extraction and Modeling for ASR .................... 97
Tue-Oral-14 Oral Session 14: Speech Analysis I ................................................... 98
Tue-Oral-16 Oral Session 16: Source Separation and Computational Auditory Scene Analysis ................................................................. 101
Tue-SP-3 Special Session 3: Speech Technologies for Ambient Assisted Living ........ 102
Tue-Poster-10 Poster Session 10: DNN for ASR ...................................................... 103
Tue-Poster-11 Poster Session 11: Speaker Recognition — General Topics ............... 105
Tue-Poster-12 Poster Session 12: Speech Processing with Multi-modalities .......... 107
Keynote 3: Lori Lamel .................................................................................... 109
Tue-Oral-17 Oral Session 17: Normalization and Discriminative Training Methods .... 110
Tue-Oral-18 Oral Session 18: Paralinguistic and Extralinguistic Information ............ 111
Tue-Oral-19 Oral Session 19: Text Processing for Speech Synthesis ....................... 112
Tue-Oral-20 Oral Session 20: Cross-language Perception and Production ................ 113
Tue-SP-4 Special Session 4: Text-Dependent Speaker Verification With Short Utterances ................................................................. 114
Tue-Poster-13 Poster Session 13: Speech and Audio Analysis .................................. 116
Tue-Poster-14 Poster Session 14: Cross-Lingual and Adaptive Language Modeling .... 117
Tue-Poster-15 Poster Session 15: Pronunciation Modeling and Learning .................. 119
Show&Tell-1 Show and Tell Session 1 .................................................................. 121
Wednesday 17 September 2014

Keynote 4: William S.-Y. Wang ........................................................... 123
Wed-Oral-23 Oral Session 23: Disordered Speech ................................................ 126
Wed-Oral-24 Oral Session 24: Speech and Multimodal Resources ........................ 127
Wed-SP-5 Special Session 5: Phase Importance in Speech Processing Applications ... 128
Wed-Poster-16 Poster Session 16: Phonetics and Phonology .............................. 130
Wed-Poster-17 Poster Session 17: Spoken Term Detection and Document Retrieval ... 132
Wed-Poster-18 Poster Session 18: Prosody and Paralinguistic Information .......... 134
Wed-Oral-26 Oral Session 26: Topic Spotting and Summarization of Spoken Documents ..... 138
Wed-Oral-27 Oral Session 27: DNN Learning .............................................. 139
Wed-SP-6(a) Special Session 6(a): Deep Neural Networks for Speech Generation and Synthesis I ... 141
Wed-Poster-19 Poster Session 19: Speech Analysis and Perception ........................ 142
Wed-Poster-20 Poster Session 20: Intelligibility Enhancement and Predictive Measures ... 145
Wed-Poster-21 Poster Session 21: Speech and Language Processing — General Topics ... 146
Show&Tell-2 Show and Tell Session 2 ...................................................... 148
Wed-Oral-29 Oral Session 29: Language, Dialect and Accent Recognition ............... 150
Wed-Oral-30 Oral Session 30: Adaptation ............................................... 151
Wed-Oral-31 Oral Session 31: Speaker Localization ......................................... 153
Wed-Oral-32 Oral Session 32: Speech Analysis II .......................................... 154
Wed-SP-6(b) Special Session 6(b): Deep Neural Networks for Speech Generation and Synthesis II ... 155
Wed-Poster-22 Poster Session 22: Speech Synthesis II ...................................... 156
Wed-Poster-23 Poster Session 23: Speech Representation, Detection and Classification ... 158
Wed-Poster-24 Poster Session 24: Feature Extraction and Modeling for ASR .......... 160

Thursday 18 September 2014

Keynote 5: Li Deng ............................................................................. 163
Thu-Oral-33 Oral Session 33: Spoken Term Detection for Low-Resource Languages I .... 163
Thu-Oral-34 Oral Session 34: Voice Conversion ............................................ 164
Thu-Oral-35 Oral Session 35: Speech and Audio Segmentation and Classification .... 166
Thu-Oral-36 Oral Session 36: Language Acquisition ...................................... 167
Thu-Oral-37 Oral Session 37: Speech Perception ......................................... 168
Thu-Poster-25 Poster Session 25: Language and Lexical Modeling ....................... 169
Thu-Poster-26 Poster Session 26: Speech Enhancement (Single- and Multi-Channel) ........ 171
Thu-Poster-27 Poster Session 27: Robust ASR .......................................... 173
Thu-Oral-38 Oral Session 38: Spoken Term Detection for Low-Resource Languages II ... 175
Thu-Oral-40 Oral Session 40: Speech Enhancement ........................................ 177
Thu-Oral-41 Oral Session 41: Unsupervised or Corrective Lexical Modeling ............ 179
Thu-Oral-42 Oral Session 42: Meta Data .................................................... 180
Thu-Poster-28 Poster Session 28: Speech Synthesis III ...................................... 181
Thu-Poster-29 Poster Session 29: Adaptation .............................................. 183
Thu-Poster-30 Poster Session 30: Language Recognition .................................... 185
Learning About Speech

Anne Cutler; University of Western Sydney, Australia

We human language users learn about speech all our lives. In fact if the beginning of our individual life is taken to be the moment of our birth, we learn for more than our whole lives, because even prior to birth we acquire much basic information about language sound structure. Then in early life, before we are capable of uttering any recognizable words, we have already learned a huge amount about the sounds and words of our language; furthermore, the efficiency with which we learn this is predictive of our later linguistic facility. We learn to process speech in the way best suited to the particular language (or languages) we acquire as children. This leads to a formidable efficiency and robustness in native-language processing, but it actually inhibits learning in the case of other languages we encounter only later — children are clearly very much better at learning a new language than adults are! Learning about speech nonetheless continues throughout life, in particular the everyday perceptual learning that enables us to adapt our speech processing to newly encountered talkers, and to adjust our own pronunciation in keeping with pronunciation changes in our speech community across time. This learning (at least in the native language) can draw on a wide variety of information sources, is fully in place in childhood, and is apparently unattenuated in older language users.

Oral Session 1: Multi-Lingual ASR

Garnet 213-218
11:00 – 13:00, Monday 15 September 2014
Chairs: Nelson Morgan and Chin-Hui Lee

Language ID-Based Training of Multilingual Stacked Bottleneck Features

Yu Zhang, Ekapol Chuangsuwanich, James R. Glass; MIT, USA

In this paper, we explore multilingual feature-level data sharing via Deep Neural Network (DNN) stacked bottleneck features. Given a set of available source languages, we apply language identification to pick the language most similar to the target language, for more efficient use of multilingual resources. Our experiments with IARPA-Babel languages show that bottleneck features trained on the most similar source language perform better than those trained on all available source languages. Further analysis suggests that only data similar to the target language is useful for multilingual training.

Kernel Density-Based Acoustic Model with Cross-Linguagal Bottleneck Features for Resource Limited LVCSR

Van Hai Do¹, Xiong Xiao², Eng Siong Chng¹, Haizhou Li¹;¹Nanyang Technological University, Singapore;¹²TL@NTU, Singapore

Conventional acoustic models, such as Gaussian mixture models (GMM) or deep neural networks (DNN), cannot be reliably estimated when there are very little speech training data, e.g. less than 1 hour. In this paper, we investigate the use of a non-parametric kernel density estimation method to predict the emission probability of HMM states. In addition, we introduce a discriminative score calibrator to improve the speech class posteriors generated by the kernel density for speech recognition task. Experimental results on the Wall Street Journal task show that the proposed acoustic model using cross-lingual bottleneck features significantly outperforms GMM and DNN models for limited training data case.

Improving ASR Performance On Non-Native Speech Using Multilingual and Crosslingual Information

Ngoc Thang Vu, Yuanfan Wang, Marten Klose, Zlatka Mihaylova, Tanja Schultz; KT, Germany

This paper presents our latest investigation of automatic speech recognition (ASR) on non-native speech. We first report on a non-native speech corpus — an extension of the GlobalPhone database — which contains English with Bulgarian, Chinese, German and Indian accent and German with Chinese accent. In this case, English is the spoken language (L2) and Bulgarian, Chinese, German and Indian are the mother tongues (L1) of the speakers. Afterwards, we investigate the effect of multilingual acoustic modeling on non-native speech. Our results reveal that a bilingual L1-L2 acoustic model significantly improves the ASR performance on non-native speech. For the case that L1 is unknown or L1 data is not available, a multilingual ASR system trained without L1 speech data consistently outperforms the monolingual L2 ASR system. Finally, we propose a method called crosslingual accent adaptation, which allows using English with Chinese accent to improve the German ASR on German with Chinese accent and vice versa. Without using any intra lingual adaptation data, we achieve 15.8% relative improvement in average over the baseline system.

Language Independent and Unsupervised Acoustic Models for Speech Recognition and Keyword Spotting

Kate M. Knill, Mark J.F. Gales, Anton Ragni, Shakti P. Rath; University of Cambridge, UK

Developing high-performance speech processing systems for low-resource languages is very challenging. One approach to address the lack of resources is to make use of data from multiple languages. A popular direction in recent years is to train a multi-language bottleneck DNN. Language dependent and/or multi-language (all training languages) Tandem acoustic models (AM) are then trained. This work considers a particular scenario where the target language is unseen in multi-language training and has limited language model training data, a limited lexicon, and acoustic training data without transcriptions. A zero acoustic resources case is first described where a multi-language AM is directly applied, as a language independent AM (LIAM), to an unseen language. Secondly, in an unsupervised approach a LIAM is used to obtain hypotheses for the target language acoustic data transcriptions which are then used in training a language dependent AM. 3 languages from the IARPA IARPA project are used for assessment: Vietnamese, Haitian Creole and Bengali. Performance of the zero acoustic resources system is found to be poor, with keyword spotting at best 60% of language dependent performance. Unsupervised language dependent training yields performance gains. For one language (Haitian Creole) the Babel target is achieved on the in-vocabulary data.

NOTES
Cross-Linguual Adaptation with Multi-Task Adaptive Networks
Peter Bell, Joris Driesen, Steve Renals; University of Edinburgh, UK

Posterior-based or bottleneck features derived from neural networks trained on out-of-domain data may be successfully applied to improve speech recognition performance when data is scarce for the target domain or language. In this paper we combine this approach with the use of a hierarchical deep neural network (DNN) network structure — which we term a multi-level adaptive network (MLAN) — and the use of multitask learning. We have applied the technique to cross-lingual speech recognition experiments on recordings of TED talks and European Parliament sessions in English (source language) and German (target language). We demonstrate that the proposed method can lead to improvements over standard methods, even when the quantity of training data for the target language is relatively high. When the complete method is applied, we achieve relative WER reductions of around 13% compared to a monolingual hybrid DNN baseline.

On Recognition of Non-Native Speech Using Probabilistic Lexical Model
Marzieh Razavi, Mathew Magimai Doss; Idiap Research Institute, Switzerland

Despite various advances in automatic speech recognition (ASR) technology, recognition of speech uttered by non-native speakers is still a challenging problem. In this paper, we investigate the role of different factors such as type of lexical model and choice of acoustic units in recognition of speech uttered by non-native speakers. More precisely, we investigate the influence of the probabilistic lexical model in the framework of Kullback-Leibler divergence based hidden Markov model (KL-HMM) approach in handling pronunciation variabilities by comparing it against hybrid HMM/artificial neural network (ANN) approach where the lexical model is deterministic. Moreover, we study the effect of acoustic units (being context-independent or clustered context-dependent phones) on ASR performance in both KL-HMM and hybrid HMM/ANN frameworks. Our experimental studies on French part of MediaParl as a bilingual corpus indicate that the probabilistic lexical modeling approach in the KL-HMM framework can capture the pronunciation variations present in non-native speech effectively. More precisely, the experimental results show that the KL-HMM system using context-dependent acoustic units and trained solely on native speech data can lead to better ASR performance than adaptation techniques such as maximum likelihood linear regression.

Oral Session 2: Prosody Processing
Peridot 202-203
11:00 – 13:00, Monday 15 September 2014
Chairs: Gérard Bailly and Jürgen Trouvain

Direct $F_0$ Control of an Electrolarynx Based on Statistical Excitation Feature Prediction and its Evaluation Through Simulation
Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura; NAIST, Japan

An electrolarynx is a device that artificially generates excitation sounds to enable laryngectomees to produce electrolaryngeal (EL) speech. Although proficient laryngectomees can produce quite intelligible EL speech, it sounds very unnatural due to the mechanical excitation produced by the device. To address this issue, we have proposed several EL speech enhancement methods using statistical voice conversion and showed that statistical prediction of excitation parameters, such as $F_0$ patterns, was essential to significantly improve naturalness of EL speech. In these methods, the original EL speech is recorded with a microphone and the enhanced EL speech is presented from a loudspeaker in real time. This framework is effective for telecommunication but it is not suitable to face-to-face conversation because both the original EL speech and the enhanced EL speech are presented to listeners. In this paper, we propose direct $F_0$ control of the electrolarynx based on statistical excitation prediction to develop an EL speech enhancement technique also effective for face-to-face conversation. $F_0$ patterns of excitation signals produced by the electrolarynx are predicted in real time from the EL speech produced by the laryngectomee’s articulation of the excitation signals with previously predicted $F_0$ values. A simulation experiment is conducted to evaluate the effectiveness of the proposed method. The experimental results demonstrate that the proposed method yields significant improvements in naturalness of EL speech while keeping its intelligibility high enough.

A Target Approximation Intonation Model for Yorùbá TTS
Daniel R. van Niekerk, Etienne Barnard; North-West University, South Africa

A complete intonation model based on quantitative target approximation is described for Yorùbá text-to-speech (TTS) synthesis. This model is evaluated analytically and perceptually and compared to a fundamental frequency ($F_0$) model using the standard HTS implementation. Analytical results suggest that the proposed approach more efficiently models $F_0$ contours given typical data constraints in under-resourced environments and perceptual results comparing the proposed model with HTS are encouraging.

Learning Continuous-Valued Word Representations for Phrase Break Prediction
Anandaswarup Vadapalli, Kishore Prabhallad; IIIT Hyderabad, India

Phrase break prediction is the first step in modeling prosody for text-to-speech systems (TTS). Traditional methods of phrase break prediction have used discrete linguistic representations (like POS tags, induced POS tags, word-terminal syllables) for modeling these breaks. However these discrete representations suffer from a number of issues such as fixing the number of discrete classes and also such a representation does not capture the co-occurrence statistics of the words. As a result, the use of continuous valued word representation was proposed in literature. In this paper, we propose a neural network dictionary learning architecture to induce task specific continuous valued word representations, and show that these task specific features perform better at phrase break prediction as compared to continuous features derived using Latent Semantic Analysis (LSA).

Improving Mandarin Prosodic Boundary Prediction with Rich Syntactic Features
Hao Che, Jianhua Tao, Ya Li; Chinese Academy of Sciences, China

Previous researches indicated that the performance of automatic prosodic boundary labeling benefited from syntactic phrase infor-
mation for Mandarin. However, the influence of other syntactic features such as dependency has not been studied in-depth yet, especially on large scale corpus. This paper demonstrates the usefulness of rich syntactic features for Mandarin phrase boundary prediction. Both syntactic phrase and dependency features are considered in our methods. The experimental results show that rich syntactic features improve the performance of prosodic boundary prediction effectively.

Investigating Automatic & Human Filled Pause Insertion for Speech Synthesis

Rasmus Dall\(^1\), Marcus Tomalin\(^2\), Mirjam Wester\(^1\), William Byrne\(^2\), Simon King\(^1\); \(^1\)University of Edinburgh, UK; \(^2\)University of Cambridge, UK

Filled pauses are pervasive in conversational speech and have been shown to serve several psychological and structural purposes. Despite this, they are seldom modelled overtly by state-of-the-art speech synthesis systems. This paper seeks to motivate the incorporation of filled pauses into speech synthesis systems by exploring their use in conversational speech, and by comparing the performance of several automatic systems inserting filled pauses into fluent text. Two initial experiments are described which seek to determine whether people’s predicted insertion points are consistent with actual practice and/or with each other. The experiments also investigate whether there are ‘right’ and ‘wrong’ places to insert filled pauses. The results show good consistency between people’s predictions of usage and their actual practice, as well as a perceptual preference for the ‘right’ placement. The third experiment contrasts the performance of several automatic systems that insert filled pauses into fluent sentences. The best performance (determined by F-score) was achieved through the by-word interpolation of probabilities predicted by Recurrent Neural Network and 4gram Language Models. The results offer insights into the use and perception of filled pauses by humans, and how automatic systems can be used to predict their locations.

The Effect of Filled Pauses and Speaking Rate on Speech Comprehension in Natural, Vocoded and Synthetic Speech

Rasmus Dall, Mirjam Wester, Martin Corley; University of Edinburgh, UK

It has been shown that in natural speech filled pauses can be beneficial to a listener. In this paper, we attempt to discover whether listeners react in a similar way to filled pauses in natural and vocoded speech compared to natural speech. We present two experiments focusing on reaction time to a target word. In the first, we replicate earlier work in natural speech, namely that listeners respond faster to a target word after a filled pause than following a silent pause. This is replicated in vocoded but not in synthetic speech. Our second experiment investigates the effect of speaking rate on reaction times as this was potentially a confounding factor in the first experiment. Evidence suggests that slower speech rates lead to slower reaction times in synthetic and in natural speech. Moreover, in synthetic speech the response to a target word after a filled pause is slower than after a silent pause. This finding, combined with an overall slower reaction time, demonstrates a shortfall in current synthesis techniques. Remedying this could help make synthesis less demanding and more pleasant for the listener, and reaction time experiments could thus provide a measure of improvement in synthesis techniques.

Introducing I-Vectors for Joint Anti-Spoofing and Speaker Verification

Elie Khoury\(^1\), Tomi Kinnunen\(^2\), Aleksandr Sizov\(^2\), Zhizheng Wu\(^3\), Sébastien Marcel\(^1\); \(^1\)Idiap Research Institute, Switzerland; \(^2\)University of Eastern Finland, Finland; \(^3\)Nanyang Technological University, Singapore

Any biometric recognizer is vulnerable to direct spoofing attacks and automatic speaker verification (ASV) is no exception; replay, synthesis and conversion attacks all provoke false acceptances unless countermeasures are used. We focus on voice conversion (VC) attacks. Most existing countermeasures use full knowledge of a particular VC system to detect spoofing. We study a potentially more universal approach involving generative modeling perspective. Specifically, we adopt standard i-vector representation and probabilistic linear discriminant analysis (PLDA) back-end for joint operation of spoofing attack detector and ASV system. As a proof of concept, we study a vocoder-mismatched ASV and VC attack detection approach on the NIST 2006 speaker recognition evaluation corpus. We report stand-alone accuracy of both the ASV and countermeasure systems as well as their combination using score fusion and joint approach. The method holds promise.

Random Projections for Large-Scale Speaker Search

Ryan Leary, Walter Andrews; Raytheon BBN Technologies, USA

This paper describes a system for indexing acoustic feature vectors for large-scale speaker search using random projections. Given one or more target feature vectors, large-scale speaker search enables returning similar vectors (in a nearest-neighbors fashion) in sublinear time. The speaker feature space is comprised of i-vectors, derived from Gaussian Mixture Model supervectors. The index and search algorithm is derived from locality sensitive hashing with novel approaches in neighboring bin approximation for improving the miss rate at specified false alarm thresholds. The distance metric for determining the similarity between vectors is the cosine distance. This approach significantly reduced the search space by 70% with minimal increase in miss rate. When combined with further dimensionality reduction, a reduction of the search space by over 90% is also possible. All experiments are based on the NIST SRE 2010 evaluation.

Analysis of I-Vector Framework for Speaker Identification in TV-Shows

Corinne Fredouille\(^1\), Delphine Charlet\(^2\); \(^1\)LIA, France; \(^2\)Orange Labs, France

Inspired from the Joint Factor Analysis, the i-vector-based analysis has become the most popular and state-of-the-art framework for the speaker verification task. Mainly applied within the NIST/SRE evaluation campaigns, many studies have been proposed to improve more and more performance of speaker verification systems. Nevertheless, while the i-vector framework has been used in other speech processing fields like language recognition, a very few
studies have been reported for the speaker identification task on TV shows. This work was done in the REPERE challenge context, focused on the people recognition task in multimodal conditions (audio, video, text) from TV show corpora. Moreover, the challenge participants are invited for providing systems for monomodal tasks, like speaker identification. The application of the i-vector framework is investigated through different points of views: (1) some of the i-vector based approaches are compared, (2) a specific i-vector extraction protocol is proposed in order to deal with widely varying amounts of training data among speaker population, (3) the joint use of both speaker diarization and identification is finally analyzed. Based on a 533 speaker dictionary, this joint system wins the monomodal speaker identification task of the 2014 REPERE challenge.

Boosting Bonsai Trees for Efficient Features Combination : Application to Speaker Role Identification

Antoine Laurent1, Nathalie Camelin2, Christian Raymond3; 1LIMSI, France; 2LIUM, France; 3IRISA, France

In this article, we tackle the problem of speaker role detection from broadcast news shows. In the literature, many proposed solutions are based on the combination of various features coming from acoustic, lexical and semantic information with a machine learning algorithm. Many previous studies mention the use of boosting over decision stumps to combine efficiently these features. In this work, we propose a modification of this state-of-the-art machine learning algorithm changing the weak learner (decision stumps) by small decision trees, denoted bonsai trees. Experiments show that using bonsai trees as weak learners for the boosting algorithm largely improves both system error rate and learning time.

Identifying Contributors in the BBC World Service Archive

Yves Raimond, Thomas Nixon; BBC, UK

In this paper we describe the speaker identification feature of the BBC World Service Archive prototype, an experiment run by BBC R&D to investigate alternative ways of publishing large radio archives. This feature relies on diarization of individual programmes, supervisor-based speaker models, crowdsourcing for speaker identities, and a fast distributed index based on Locality Sensitive Hashing techniques to propagate these identities. We also describe how crowdsourced data can be used to continuously evaluate and refine our mapping from speaker models to speaker identities. We believe this experiment is one of the largest of its kind.

Effect of Long-Term Ageing on i-Vector Speaker Verification

Finnian Kelly1, Rahim Saeidi2, Naomi Harte1, David A. van Leeuwen1; 1Trinity College Dublin, Ireland; 2University of Eastern Finland, Finland; 3Radboud Universiteit Nijmegen, The Netherlands

Assessing the impact of ageing on biometric systems is an important challenge. In this paper, an i-vector speaker verification framework is used to evaluate the impact of long-term ageing on state-of-the-art speaker verification. Using the Trinity College Dublin Speaker Ageing (TCDSA) database, it is observed that the performance of the i-vector system, in terms of both discrimination and calibration, degrades progressively as the absolute age difference between training and testing samples increases. In the case of male speakers, the equal error rate (EER) increases from 4.61% at an ageing difference of 0–1 years to 32.74% at an age difference of 51–60 years. The performance of a Gaussian Mixture Model - Universal Background Model (GMM-UBM) system is presented for comparison. It is shown that while the i-vector system outperforms the GMM-UBM system, as absolute age difference increases, the performance of both degrades at a similar rate. It is concluded that long-term ageing variability is distinct from everyday inter-sessional variability, and therefore must be dealt with via dedicated compensation strategies.

Oral Session 4: Phonetics and Phonology

Peridot 201
11:00 – 13:00, Monday 15 September 2014
Chairs: Chi-yu Tseng and Yoshinori Sagisaka

Acoustic Correlates of Phonological Status

Maarten Versteegh1, Amanda Seidl2, Alejandrina Cristia1; 1LSCP, France; 2Purdue University, USA

Languages vary not only in terms of their sound inventory, but also in the phonological status certain sound distinctions are assigned. For example, while vowel nasality is lexically contrastive (phonemic) in Quebecois French, it is largely determined by the context (allophonic) in American English; the reverse is true for vowel tenseness. If phonetics and phonology interact, a minimal pair of sounds should span a larger acoustic divergence when it is pronounced by speakers for whom the underlying distinction is phonemic compared to allophonic. Near minimal pairs were segmented from a corpus of American English and Quebecois French using a crossed design (since nasality and tenseness have opposite phonological status in the two languages). Pairwise time-aligned divergences between contrasts were calculated on the basis of 7 mainstream spoken feature representations, and a set of linguistic phonetic measurements. Only carefully selected phonetic measurements revealed the expected cross-over, with larger divergences for English than French tokens of the tenseness contrast, and larger divergences for French than English tokens for the nasality contrast. We conclude that the phonetic effects of phonological status are subtle enough that only linguistically-informed (or supervised) measurements can pick up on them.

Parameterization of the Glottal Source with the Phase Plane Plot

Manu Airaksinen, Paavo Alku; Aalto University, Finland

Parameterization of the glottal flow is a process where the glottal flow is represented in terms of a few numerical values. This study proposes a novel parameterization technique called the phase plane symmetry (PPS) parameter that utilizes the symmetrical properties of the phase plane plot. Phase plane is a way to graphically visualize the glottal source in a 2-dimensional space spanned by two amplitude-domain axes. A correctly normalized phase plane plot has also close ties to the normalized amplitude quotient (NAQ) parameter, and it is shown that the inverse NAQ value is represented as a single point in the phase plane plot. The experiments conducted in this study support that PPS is powerful in discriminating between various phonation types and within the same range of robustness as the NAQ parameter.

NOTES
The accuracy of the widely used and International Phonetic Association-sanctioned Chao five-point scale of tonal transcription is examined quantitatively. Perceptually transformed acoustic data are used from two Chinese dialects with complex tone systems, and a measure derived of the conformability of the data using their likelihoods. It is shown that some tones conform well to the model, but others do not, with tonal pitch targets lying equidistant between the Chao integers. It is concluded that the Chao model is probably not an accurate reflection of the distribution of tonal pitch targets.

Intonational Phonology and Prosodic Hierarchy in Malay

Diyana Hamzah, James Sneed German; Nanyang Technological University, Singapore

This paper presents original data in support of a new model of intonational phonology for Malay as spoken in Singapore. Building on the Autosegmental-Metrical approach (Beckman & Pierrehumbert, 1986), we propose that intonational variation in Malay can be explained in terms of underlying sequences of abstract tonal units (H and l), which are aligned to the edges and internal syllables of prosodic phrases organized in a hierarchy. Data was drawn from a production experiment (Hamzah, 2012) involving declarative utterances under different focus patterns in a question-answer context, as well as from story-telling interviews and TV interviews. We find evidence for at least three levels of prosodic organization: (i) an accentual phrase which comprises one or more words and bears an L and H tone at its left and right edges, respectively, (ii) an intermediate phrase, which serves as the domain of catapheresis, and (iii) an intonational phrase, which may span the entire utterance and bears an additional H or L tone at its right edge. Differences in F0 peak alignment for focused words support the presence of a focus pitch accent. We outline a series of follow-up studies for extending the model further.

Comparing Parameterizations of Pitch Register and its Discontinuities at Prosodic Boundaries for Hungarian

Uwe D. Reichel 1, Katalin Mády 2; 1LMU München, Germany; 2Hungarian Academy of Sciences, Hungary

We examined how well prosodic boundary strength can be captured by two declination stylization methods as well as by four different representations of pitch register. In the stylization proposed by Liebermann et al. (1985) base- and topline are fitted to peaks and valleys of the pitch contour, whereas in Reichel & Mády (2013) these lines are fitted to medians below and above certain pitch percentiles. From each of the stylizations four feature pools were used representing different aspects of register discontinuity at word boundaries: discontinuities related to the base, mid, and topline, as well as to the range between base- and topline. Concerning stylization the median-based fitting approach turned out to be more robust with respect to declination line crossing errors and yielded base-, topline and range-related discontinuity characteristics with higher correlations to perceived boundary strength. Concerning register representation, for the peak-valley fitting approach the base- and topline patterns showed weaker correspondences to boundary strength than the other feature pools. We furthermore trained generalized linear regression models for boundary strength prediction on each feature pool. It turned out that neither the stylization method nor the register representation had a significant influence on the overall good prediction performance.

An Evaluation of Machine Learning Methods for Prominence Detection in French

George Christodoulides 1, Mathieu Avanzi 2; 1Université catholique de Louvain, Belgium; 2LLF (UMR 7110), France

The automatic detection of prosodically prominent syllables is crucial for analysing speech, especially in French where prominence contributes substantially to prosodic grouping and boundary demarcation. In this paper, we compare different machine learning techniques for the automatic detection of prominent syllables, using prosodic features (including pitch, energy, duration and spectral balance) and lexical information. We explore the differences between modelling the detection of prominent syllables as a classification or as a sequence labelling problem, and combinations of the two techniques. We train and evaluate our systems on a corpus of spontaneous French speech, consisting of almost 100 different speakers; the corpus is balanced for speaker age and sex and covers 3 different regional varieties. The result of this study is a novel tool for the automatic annotation of prominent syllables in French.

Learning Situated Knowledge Bases Through Dialog

Aasish Pappu, Alexander I. Rudnicky; Carnegie Mellon University, USA

To respond to a user’s query, dialog agents can use a knowledge base that is either domain specific, commonsense (e.g., NELL, Freebase) or a combination of both. The drawback is that domain-specific knowledge bases will likely be limited and static; commonsense ones are dynamic but contain general information found on the web and will be sparse with respect to a domain. We address this issue through a system that solicits situational information from its users in a domain that provides information on events (seminar talks) to augment its knowledge base (covering an academic field). We find that this knowledge is consistent and useful and that it provides reliable information to users. We show that, in comparison to a base system, users find that retrievals are more relevant when the system uses its informally acquired knowledge to augment their queries.

Crowdsourcing for Situated Dialog Systems in a Moving Car

Teruhisa Misu; Honda Research Institute USA, USA

In this paper, we address issues that arise when crowdsourcing data collection of user queries to situated dialog systems in a moving car. Compared to unimodal spoken dialog systems such as systems for smartphones, collecting dialog data for situated dialog
systems is more costly because a clear awareness of the physical surroundings is required for the user to make realistic queries. We consider the use of crowdsourcing to collect them. To elicit queries from crowd workers, we propose methods of prompting them using visual information. The queries collected using the crowdsourcing methods are compared to those collected using a real situated dialog system. Specifically, we evaluate them based on several performance measures of similarity in semantic content, naturalness of language expression and bias of the collected data. We demonstrate that our crowdsourcing method produced a better language resource in terms of the similarity of the text to real user utterances than those generated by a handicraft grammar.

Evaluating Coherence in Open Domain Conversational Systems
Ryuichiro Higashinaka, Toyomi Meguro, Kenji Imamura, Hiroaki Sugiyama, Toshiro Makino, Yoshihiro Matsuo; NTT Corporation, Japan

We propose a method for evaluating coherence between user utterances and those generated from open-domain conversational systems. Our aim is to make it possible for such systems to ascertain whether utterances generated from them are appropriate to the context before generation so that possible breakdown in conversation arising from inappropriate utterances can be avoided. In our method, we train a classifier that distinguishes a pair of a user utterance and that generated from a system as coherent or incoherent by using various pieces of information related to dialogue exchange, such as dialogue acts, question types, and predicate-argument structures. Experimental results show that our method significantly outperforms the baseline, confirming its effectiveness.

Adapting Dependency Parsing to Spontaneous Speech for Open Domain Spoken Language Understanding
Frederic Bechet, Alexis Nasr, Benoît Favre; LIF (UMR 7279), France

Parsing human-human conversations consists in automatically enriching text transcription with semantic structure information. We use in this paper a FrameNet-based approach to semantics that, without needing a full semantic parse of a message, goes further than a simple flat translation of a message into basic concepts. FrameNet-based semantic parsing may follow a syntactic parsing step, however spoken conversations in customer service telephone call centers present very specific characteristics such as non-canonical language, noisy messages (disfluencies, repetitions, truncated words or automatic speech transcription errors) and the presence of superfluous information. For syntactic parsing the traditional view based on context-free grammars is not suitable for processing non-canonical text. New approaches to parsing based on dependency structures and discriminative machine learning techniques are more adapted to process spontaneous speech for two main reasons: (a) they need less training data and (b) the annotation with syntactic dependencies of conversation transcripts is simpler than with syntactic constituents. Another advantage is that partial annotation can be performed. This paper presents the adaptation of a syntactic dependency parser to process very spontaneous speech recorded in a call-centre environment. This parser is used in order to produce FrameNet candidates for characterizing conversations between an operator and a caller.

Incremental On-Line Adaptation of POMDP-Based Dialogue Managers to Extended Domains
M. Gašić, Dongho Kim, Pirros Tsiakoulis, Catherine Breslin, Matthew Henderson, M. Szummer, B. Thomson, Steve Young; University of Cambridge, UK

An important property of open domain spoken dialogue systems is their ability to deal with a set of new, previously unseen, concepts introduced in the conversation. The dialogue manager must then quickly learn how to talk about the new concepts using its knowledge of the existing concepts. It has previously been shown that a single new concept could be accommodated by mapping the kernel function of a Gaussian process to incorporate an additional concept into the domain of a statistical dialogue manager. Here we present an incremental scheme which enables the domain of a dialogue manager to be repeatedly extended by recursively specifying priors in Gaussian processes. We show that it is possible to effectively double the number of concepts understood by a system providing restaurant information using only 1000 adaptation dialogues with real users.

Hypotheses Ranking for Robust Domain Classification and Tracking in Dialogue Systems
Jean-Philippe Robichaud, Paul A. Crook, Puyang Xu, Omar Zia Khan, Ruhi Sarikaya; Microsoft, USA

We present a novel application of hypothesis ranking (HR) for the task of domain detection in a multi-domain, multi-turn dialog system. Alternate, domain dependent, semantic frames from a spoken language understanding (SLU) analysis are ranked using a gradient boosted decision trees (GBDT) ranker to determine the most likely domain. The ranker, trained using Lambda Rank, makes use of a range of signals derived from the SLU and previous turn context to improve domain detection. On a multi-turn corpus we show that this approach offers accuracy improvements of 3.2% absolute (25.6% relative) compared to relying solely on upfront non-contextual SLU domain models and 2.9% (24.5% relative) improvement even with contextual SLU domain models. We also show that HR can be trained to be robust to changes in the SLU.

Motor Control Primitives Arising from a Learned Dynamical Systems Model of Speech Articulation
Vikram Ramanarayanan, Louis Goldstein, Shrikanth S. Narayanan; University of Southern California, USA

We present a method to derive a small number of speech motor control "primitives" that can produce linguistically-interpretable articulatory movements. We envision that such a dictionary of primitives can be useful for speech motor control, particularly in finding a low-dimensional subspace for such control. First, we use the iterative Linear Quadratic Gaussian with Learned Dynamics (ILOG-LD) algorithm to derive (for a set of utterances) a set of stochastically optimal control inputs to a learned dynamical systems model of the vocal tract that produces desired movement sequences. Second, we use a convolutive Nonnegative Matrix Fac-
torization with sparseness constraints (cNMFsc) algorithm to find a small dictionary of control input primitives that can be used to reproduce the aforementioned optimal control inputs that produce the observed articulatory movements. The method performs favorably on both qualitative and quantitative evaluations conducted on synthetic data produced by an articulatory synthesizer. Such a primitives-based framework could help inform theories of speech motor control and coordination.

**Nonword Repetition of Taiwanese Disyllabic Tonal Sequences in Adults with Language Attrition**

Chia-Hsin Yeh$^1$, Chiung-Yao Wang$^2$, Jung-Yueh Tu$^3$; $^1$Michigan State University, USA; $^2$University of Colorado at Boulder, USA; $^3$National Taiwan Normal University, Taiwan

Mon-P-1-Z, Poster

This study demonstrates the nonword repetition format comparable to other conventional tasks (picture-naming, reading, and so on) as a plausible measure of linguistic competence for adults with language attrition. Taiwanese speakers with and without attrition symptoms, defined by frequency of use, were recruited, and so were American learners of Mandarin Chinese. The results show that (1) fluent speakers’ repetition accuracy of Taiwanese tones is significantly higher than attrition speakers’, and American learners’ is the worst, (2) among five target tones (high-level, low-rising, low-falling, high-falling and mid-level), the repetition accuracy of high-falling tone is the highest, and that of low-level tone is the lowest in non-word-final position across the three participant groups, and (3) the least accurate mid-level tone tends to be mispronounced as low-rising. The findings suggest that the participants’ frequency of use and exposure to Taiwanese is positively correlated with the repetition accuracy, and mid-level tone is the most difficult category to learn. More crucially, the percent accuracy and confusion matrix of nonword repetition enlighten how midlevel tone is more susceptible to sound change.

**A Unified Account of Prominence Effects in an Optimization-Based Model of Speech Timing**

Andreas Windmann$^1$, Juraj Simko$^2$, Petra Wagner$^1$; $^1$Universität Bielefeld, Germany; $^2$University of Helsinki, Finland

Mon-P-1-3, Poster

We show how our optimization-based model of speech timing reproduces three effects of prosodic prominence on suprasegmental timing patterns in speech: (1) the durational interaction between lexical stress and pitch accent, (2) polysyllabic shortening in pitch-accented words and (3), differential behavior of prominent and non-prominent syllables under speaking rate variation. We review the literature and present model simulations that replicate reported phenomena. Results underline the capacity of our model to provide a unified account of the temporal organization of speech.

**Estimation of the Movement Trajectories of Non-Crucial Articulators Based on the Detection of Crucial Moments and Physiological Constraints**

Jangwon Kim, Sungbok Lee, Shrikanth S. Narayanan; University of Southern California, USA

Mon-P-1-4, Poster

This study develops a mathematical model that estimates the movements of (linguistically) non-crucial articulators in speech production, which provides a systematic way to study the relationship between the behaviors of crucial and non-crucial articulators; crucial articulators are those essential for realizing a speech task. The underlying assumption of our model is that non-crucial articulatory movements are governed by the physiological constraints in relation to the corresponding crucial articulators as well as by the contextual constraint from the nearest crucial time of the non-crucial articulator. These constraints have been generally assumed in the speech production literature, but they have not been incorporated directly into articulatory models. The crucial articulatory moments in an utterance are automatically determined by a novel forced-alignment algorithm for articulatory trajectories, which uses the inherent physical properties of crucial articulatory movements. Experimental results suggest that the proposed algorithm is capable of estimating non-crucial articulatory positions well in both neutral and emotional speech, significantly better than the simple interpolation of crucial points.

**Sparse Smoothing of Articulatory Features from Gaussian Mixture Model Based Acoustic-to-Articulatory Inversion: Benefit to Speech Recognition**

Prasad Sudhakar$^1$, Prasanta Kumar Ghosh$^2$; $^1$Université catholique de Louvain, Belgium; $^2$Indian Institute of Science, India

Mon-P-1-5, Poster

Speech recognition using articulatory features estimated using Acoustic-to-Articulatory Inversion (AAI) is considered. A recently proposed sparse smoothing approach is used to postprocess the estimates from Gaussian Mixture Model (GMM) based AAI using Minimum Mean Squared Error (MMSE) criterion. It is well known that low-pass smoothing as post-processing improves the AAI performance. Sparse smoothing, on the other hand, not only improves the AAI performance but also preserves the MMSE optimality for as many estimates as possible. In this work we investigate the benefit of preserving MMSE optimality during postprocessing by using the smoothed articulatory estimates in a broad class phonetic recognition task. Experimental results show that the low-pass filter based smoothing results in a significant drop in the recognition accuracy compared to that using articulatory estimates without any smoothing. However, the recognition accuracy obtained by articulatory features from sparse smoothing is similar to that using articulatory features directly from GMM based AAI without any postprocessing. Thus, sparse smoothing provides benefit both in terms of the inversion performance as well as recognition accuracy, while that is not the case with low-pass smoothing.

**Contribution of Tongue Lateral to Consonant Production**

Jun Wang, William Katz, Thomas F. Campbell; University of Texas at Dallas, USA

Mon-P-1-6, Poster

Speaking requires the coordinated movements of individual articulators. Understanding each articulator’s contribution to speech is fundamental not only for understanding how speech is produced, but also for optimizing speech assessment and treatment. Our recent work has studied the individual contributions of tongue tip, tongue blade, tongue body front, tongue body back, upper lip, and lower lip movement to speech sound production by tracking the motion of sensors attached on the midline of tongue and lips. An optimal set of articulators (tongue tip, tongue body back, upper lip, and lower lip) has been found. However, the tongue lateral (side)’s contribution to speech is still poorly understood.

**Notes**
We therefore investigated the contribution of the tongue lateral region to consonant production by analyzing the motion of a sensor attached to the side of tongue. Repeated productions of 12 consonants (including the lateral approximant /l/) were collected from six native English speakers. Consonant classification accuracy based on articulatory movement data obtained using a support vector machine was used as an indication of contribution level. The results suggest that sagittal movement of the tongue lateral sensor did not significantly benefit consonant classification, over and above the optimal set. Implications of these findings are discussed.

A Preliminary Study on Acoustic Correlates of Tone2+Tone2 Disyllabic Word Stress in Mandarin

Min Liu¹, Shuju Shi², Jinsong Zhang²;¹Universiteit Leiden, The Netherlands;²BLCU, China

This paper investigated the potential acoustic correlates of word stress within a disyllabic tonal sequence, a rising tone followed by a rising tone (Tone2 + Tone2) in Mandarin, based on a large corpus with adequate information of stress patterns and prosodic boundary levels. The results showed that a) For Tone2+Tone2 words, features based on tone nucleus were more effective than that of the whole F0 contour for stress identification. Particularly, three new acoustic correlates of stress were proposed, namely, the F0 change difference, the duration difference and the F0 slope difference of tone nucleus segment between the two syllables. b) These three parameters could serve properly as acoustic cues to differentiate initial-stressed and final-stressed words, as well as initial-stressed and equal-stressed words irrespective of prosodic boundary levels. c) With the increase of prosodic boundary levels, the duration difference of the tone nucleus segment got smaller for the initial-stressed words while getting larger for the final-stressed ones. d) At higher prosodic boundary words, due to stronger enlargement of pre-boundary lengthening at the initial-stressed words while getting larger for the final-stressed ones, the duration difference of the tone nucleus segment got smaller for the higher prosodic boundary levels. e) The compensation effect between the higher prosodic boundary levels. The results showed that a) For Tone2+Tone2 words, features based on tone nucleus were more effective than that of the whole F0 contour for stress identification. Particularly, three new acoustic correlates of stress were proposed, namely, the F0 change difference, the duration difference and the F0 slope difference of tone nucleus segment between the two syllables. 

Vowel Length Impact on Locus Equation Parameters: An Investigation on Jordanian Arabic

Mohammad Abuoudeh, Olivier Crouzet; LLING (EA3827), France

Investigations of changes in speech rate and speaking style have reported that locus equation parameters are sensible to temporal variation though they usually do not evidence any major impact on place of articulation classification. The aim of the present study was to investigate the role of vowel length contrasts on locus equation parameters in Jordanian Arabic (JA) speech production in order to study the influence of intrinsic temporal changes (vowel length contrasts) on locus equation parameters. Statistical analyses of slopes and intercepts show that the locus equation parameters of a consonant produced with long vowels are significantly different from those of the same consonant produced with short vowels. Though this observation confirms the impact of time variations on locus equations, further analyses will need to address issues related to the potential impact of duration-related spectral modifications and the role these changes may play on the perceptual categorization of consonantal place of articulation.

Corpus-Testing a Fricative Discriminator; or, Just How Invariant is this Invariant?

Philip J. Roberts¹, Henning Reetz², Aditi Lahiri¹;¹University of Oxford, UK;²Goethe-Universität Frankfurt am Main, Germany

Acoustic cues to the distinction between bilabial fricatives are claimed to be invariant across languages. In [1], Evers et al. present a method for distinguishing automatically between [s] and [ʃ], using the slope of regression lines over separate frequency ranges within a DFT spectrum. They report accuracy rates in excess of 98% for fricatives extracted from minimal pairs in English, Dutch and Bengali. These findings are broadly replicated by [2], using VCV tokens recorded in the lab. We tested the algorithm from [1] against tokens of fricatives extracted from the TIMIT corpus of American English read speech, and the Kiel corpora of German. We were able to achieve similar accuracy rates to those reported in [1] and [2], with the following caveats: (1) the measure relies on being able to perform a DFT for frequencies from 0 to 8 kHz, so that a minimum sampling rate of 16 kHz is necessary for it to be effective, and (2) although the measure draws a similarly clear distinction between [s] and [ʃ] to those found in previous studies, the absolute value of the threshold between the two sounds is sensitive to the dynamic range of the input signal.

Modeling Coarticulation in Continuous Speech

Brian O. Bush, Alexander Kain; Oregon Health & Science University, USA

Modeling coarticulation in speech has been largely limited to short sequences and/or limited phonetic context. We introduce a methodology for modeling both formant frequency and bandwidth in continuous speech, allowing examination of sentence-level coarticulation. The model represents continuous trajectories as a combination of overlapping local trajectories, which are represented by a weighted-addition of acoustic event targets by sigmoidal coarticulation functions characterized by slope and position. Estimation is achieved using a combination of hill-climbing and grid-search, with global target, joint slope for identical contexts, and local position parameters. We evaluate model performance for two speakers using an intelligibility test that compares vocoded model output to a purely vocoded and a natural condition.

On Classification Between Normal and Pathological Voices Using the MEEI-KayPENTAX Database: Issues and Consequences

Khalid Daoudi¹, Blaise Bertrac²;¹INRIA, France;²Université Bordeaux 1, France

A large amount of research in pathological voice classification consider the task of feature extraction for discrimination between normal and dysphonic sustained vowels. The most widely used dataset for this purpose is the Massachusetts Eye & Ear Infirmary (MEEI) Voice Disorders Database commercialized by KayPENTAX Corp. During the last two decades, dozens of methods have been proposed to extract discriminative features from these signals in order to design accurate classifiers between the two classes of this database. The main contribution of this paper is to show that the normal and dysphonic sustained vowels of the KayPENTAX database are actually perfectly separable. This implies that this

NOTES
dataset is not suited for the normal-vs-dysphonic classification task, as long as the only concern is to achieve high classification accuracy. Indeed, we show that a single scalar parameter extracted from a matching pursuit decomposition of these signals (with a Gabor dictionary) yields a perfect classification accuracy (100% with a large margin). We then discuss the implication of this finding on the precaution that should be taken with this database and on research in pathological voice detection in general.

Synchonic Variation in the Articulation and the Acoustics of the Polish Three-Way Place Distinction in Sibilants and its Implications for Diachronic Change

Véronique Bukmaier, Jonathan Harrington, Ulrich Reubold, Felicitas Kleber; LMU München, Germany

The aim of the present study was to relate articulatory properties of the Polish sibilants /s/ and /ś/, the former having occurred in a number of Polish dialects. For this purpose tongue tip (TT) movement data was obtained together with acoustic data using electromagnetic articolography. The sibilants, that were always followed by either /a/ or /o/, were produced by four L1-Polish speakers at fast and slow speech rates. While /s/ had almost identical transitions, they differed greatly in the spectral characteristics with /ś/ being closer to /ś/. In order to capture differences in tongue position as well as shape both TT position and TT orientation data were analyzed. The vertical TT orientation showed similarities in /s/ and /ś/ production, but the two sibilants were clearly separated in TT position, with /ś/ being produced far more back than /s/ and /ś/, and the latter two being very similar. The tendentially greater effect of speech rate on /s/ together with the varying acoustic and articulatory similarities between the sibilants are taken as an indicator for greater instability of /ś/. This synchronic instability is discussed in terms of potential diachronic mergers.

Poster Session 2: Extraction of Para-Linguistic Information

Max Atria Gallery
11:00 – 13:00, Monday 15 September 2014
Chair: Omid Sadjadi

Predicting Client’s Inclination Towards Target Behavior Change in Motivational Interviewing and Investigating the Role of Laughter

Rahul Gupta 1, Panayiotis G. Georgiou 1, David C. Atkins 2, Shrikant S. Narayanan 1; 1University of Southern California, USA; 2University of Washington, USA

Motivational interviewing (MI) is a goal oriented psychotherapy that facilitates intrinsic motivation within a client in order to change behavior in a dialog setting. The Motivational Interviewing Skills Code (MISC) is a manual observational coding method used to quantify and evaluate the quality of MI sessions using their audio-visual recordings. However, this coding method is both labor intensive and expensive. We present an approach towards automating MISC assignments in MI involving addiction cure. Specifically, we focus on predicting valence for “Client Change Talk” (ChangeTalk) utterances, which indicate a client’s attitude towards a “Target Behavior Change” (Target). We further study the effect of incorporating counselor behavior in the model. We observe that our best model achieves an unweighted accuracy of 50.8% in a 3-way classification of positive vs negative valence ChangeTalk vs no ChangeTalk. Furthermore, we study the effect of including non-verbal behavior, specifically laughters, in our model. Information regarding location of laughters improves the unweighted accuracy of our model to 51.4% and our experimental results suggest prosodic differences in laughters belonging to ChangeTalk utterances with different valences.

Modeling Therapist Empathy Through Prosody in Drug Addiction Counseling

Bo Xiao 1, Daniel Bone 1, Maarten Van Segbroeck 1, Zac E. Imel 2, David C. Atkins 3, Panayiotis G. Georgiou 1, Shrikant S. Narayanan 1; 1University of Southern California, USA; 2University of Utah; 3University of Washington, USA

Empathy measures the capacity of the therapist to experience the same cognitive and emotional dispositions as the patient, and is a key quality factor in counseling. In this work we build computational models to infer the empathy of therapist using prosodic cues. We extract pitch, energy, jitter, shimmer and utterance duration from the speech signal, and normalize and quantize these features in order to estimate the distribution of certain prosodic patterns during each interaction. We find significant correlation between empathy and the distribution of prosodic patterns, and achieve 75% accuracy in classifying therapist empathy levels using this distribution. Experiment results suggest high pitch and energy of the therapist are negatively correlated with empathy. These observations agree with domain literature and human intuition.

An Investigation of Vocal Arousal Dynamics in Child-Psychologist Interactions Using Synchrony Measures and a Conversation-Based Model

Daniel Bone 1, Chi-Chun Lee 2, Alexandros Potamianos 1, Shrikant S. Narayanan 1; 1University of Southern California, USA; 2National Tsing Hua University, Taiwan

Researchers from various disciplines are concerned with the study of affective phenomena, especially arousal. Expressed affective modulations, which reflect both an individual’s internal state and external factors, are central to the communicative process. Bone et al. developed a robust, unsupervised (rule-based) method which provides a scale-continuous, bounded arousal rating from the vocal signal. In this study, we investigate the joint-dynamics of child and psychologist vocal arousal in autism spectrum disorder (ASD) diagnostic interactions. Arousal synchrony is assessed with multiple methods. Results indicate that children with higher ASD severity tend to lead the arousal dynamics more, seemingly because the children aren’t as responsive to the psychologist’s affective modulations. A vocal arousal model is also proposed which incorporates social and conversational constructs. The model captures conversational signal relations, and is able to distinguish between high and low ASD severity at accuracies well-above chance.
Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine
Kun Han¹, Dong Yu², Ivan Tashev²; ¹Ohio State University, USA; ²Microsoft, USA
Mon-P-2-5, Poster
Speech emotion recognition is a challenging problem partly because it is unclear what features are effective for the task. In this paper we propose to utilize deep neural networks (DNNs) to extract high level features from raw data and show that they are effective for speech emotion recognition. We first produce an emotion state probability distribution for each speech segment using DNNs. We then construct utterance-level features from segment-level probability distributions. These utterance-level features are then fed into an extreme learning machine (ELM), a simple and efficient single-hidden-layer neural network, to identify utterance-level emotions. The experimental results demonstrate that the proposed approach effectively learns emotional information from low-level features and leads to 20% relative accuracy improvement compared to the state-of-the-art approaches.

An Annotation Scheme for Sighs in Spontaneous Dialogue
Khiet P. Truong, Gerben J. Westerhof, Franciska de Jong, Dirk Heylen; University of Twente, The Netherlands
Mon-P-2-8, Poster
Sighs are non-verbal vocalisations that can carry important information about a speaker’s emotional (and psychological) state. Although sighs are commonly associated with negative emotions (e.g. giving up on something, ‘a sigh of despair’, sadness), sighs can also be associated with positive emotions such as relief. In order to gain a better understanding of sighing as a social and affective signal in dialogue, and to advance towards an automatic classification and interpretation of the emotional content of sighs, it is necessary to learn more about the various phonetic characteristics of sighs. To that end, we developed an annotation scheme for sighs that takes the variation in phonetic form into account. Using this scheme, an oral history corpus containing emotionally-coloured dialogues was annotated for sighs. Results show that sighs can be annotated with a sufficient level of reliability (Cohen’s Kappa of 0.713), and that indeed, various types of sighs can be identified as well (Cohen’s Kappa between 0.637 and 0.805). Through a preliminary analysis of emotional content words, indications were found that certain types of sighs can be associated with specific emotional contexts.

Speaker Idiosyncratic Variability of Intensity Across Syllables
Lei He, Volker Dettwo; Universität Zürich, Switzerland
Mon-P-2-6, Poster
This study explored speaker idiosyncrasy by measuring the syllabic intensity variability in the speech signal. Sixteen speakers of the TEOVOID corpus, each producing 256 read sentences, were analyzed. Characteristics of intensity variability (average or peak) between syllables were measured either holistically (standard deviation of intensity changes between syllables) or locally (pairwise variability indices of intensity changes between syllables). The results indicated significant effects of the speakers in all the metrics, suggesting a potential application of the methods for speaker recognition, and in particular for forensic speaker comparison.

Building a Naturalistic Emotional Speech Corpus by Retrieving Expressive Behaviors from Existing Speech Corpora
Soroosh Maroiooryad, Reza Lotfian, Carlos Busso; University of Texas at Dallas, USA
Mon-P-2-7, Poster
A key element in affective computing is to have large corpora of genuine emotional samples collected during natural conversations. Recording natural interactions through telephone is an appealing approach to build emotional databases. However, collecting real conversational data with expressive reactions is a challenging task, especially if the recordings are to be shared with the community (e.g., privacy concerns). This study explores a novel approach consisting in retrieving emotional reactions from existing spontaneous speech databases collected for general speech processing problems. Although most of the recordings in these databases are expected to have non-emotional expressions, given the naturalness of the interactions, the flow of the conversation can lead to emotional responses from conversation partners which we aim to retrieve. We use the IEMOCAP and SEMAINE databases to build emotion detector systems. We use these classifiers to identify emotional behaviors from the FIsher database, which is a large conversational speech corpus recorded over the phone. Subjective evaluations over the retrieved samples demonstrate the potential of the proposed scheme to build naturalistic emotional speech database.

Identification of Age-Group from Children’s Speech by Computers and Humans
Saeid Safavi, Martin Russell, Peter Jančovič; University of Birmingham, UK
Mon-P-2-8, Poster
This paper presents results on age-group identification (Age-ID) for children’s speech, using the OGI Kids corpus and GMM-UBM, GMM-SVM and i-vector systems. Regions of the spectrum containing important age information for children are identified by conducting Age-ID experiments over 21 frequency sub-bands. Results show that the frequencies above 5.5 kHz are least useful for Age-ID. The effect of using gender-independent and gender-dependent age-group modelling is explored. The GMM-UBM and i-vector systems considerably outperform the GMM-SVM system. The best Age-ID performance of 85.77% is obtained by the i-vector system applied to band-limited speech to 5.5 kHz. Experiments on human Age-ID were also conducted and the results show that the humans do not achieve the performance of the machine.

Poster Session 3: Spoken Language Understanding
Max Atria Gallery
11:00 – 13:00, Monday 15 September 2014
Chair: Ruhi Sarikaya

Theme Identification in Human-Human Conversations with Features from Specific Speaker Type Hidden Spaces
Mohamed Mochid, Richard Dufour, Mohamed Bouallegue, Georges Linarès, Renato De Mori; LIA, France
Mon-P-3-1, Poster
This paper describes a research on topic identification in a real-world customer service telephone conversations between an
agent and a customer. Separate hidden spaces are considered for agents, customers and the combination of them. The purpose is to separate semantic constituents from the speaker types and their possible relations. Probabilities of hidden topic features are then used by separate Gaussian classifiers to compute theme probabilities for each speaker type. A simple strategy, that does not require any additional parameter estimation, is introduced to classify themes with confidence indicators for each theme hypothesis. Experimental results on a real-life application show that the use of features from speaker type specific hidden spaces capture useful semantic contents with significantly superior performance with respect to independent word-based features or a single set of features. Experimental results also show that the proposed strategy makes it possible to perform surveys on collections of conversations by automatically selecting processed samples with high theme identification accuracy.

Learning Phrase Patterns for Text Classification Using a Knowledge Graph and Unlabeled Data
Alex Marin 1, Roman Holenstein 2, Ruhi Sarikaya 2, Mari Ostendorf 1, 1University of Washington, USA; 2Microsoft, USA
Mon-P-2, Poster
This paper explores a novel method for learning phrase pattern features for text classification, employing a mapping of selected words into a knowledge graph and self-training over unlabeled data. Using Support Vector Machine classification, we obtain improvements over lexical and fully-supervised phrase pattern features in domain and intent detection for language understanding, particularly in conjunction with the use of unlabeled data. Our best results are obtained using unlabeled data filtered for both model training and feature learning based on the confidence of the baseline classifiers.

Targeted Feature Dropout for Robust Slot Filling in Natural Language Understanding
Puyang Xu, Ruhi Sarikaya; Microsoft, USA
Mon-P-3, Poster
In slot filling with conditional random field (CRF), the strong current word and dictionary features tend to swamp the effect of contextual features, a phenomenon also known as feature under-training. This is a dangerous tradeoff especially when training data is small and dictionaries are limited in their coverage of the entities observed during testing. In this paper, we propose a simple and effective solution that extends the feature dropout algorithm, directly aiming at boosting the contribution from entity context. We show with extensive experiments that the proposed technique can significantly improve the robustness against unseen entities, without degrading performance on entities that are either seen or exist in the dictionary.

Spoken Question Answering Using Tree-Structured Conditional Random Fields and Two-Layer Random Walk
Sz-Rung Shiang, Hung-yi Lee, Lin-shan Lee; National Taiwan University, Taiwan
Mon-P-3-4, Poster
In this paper, we consider a spoken question answering (QA) task, in which the questions are in form of speech, while the knowledge source for answers are the webpages (in text) over the Internet to be accessed by an information retrieval engine, and we mainly focus on query formulation and re-ranking part. Because the recognition results for the spoken questions are less reliable, we use N-best lists in order to have higher probabilities to induce more correct keywords for the questions, but more noisy words are inevitably included as well. We therefore propose a hierarchical labeling method using tree-structured conditional random fields (CRF) to leverage the parse tree information or the syntactic structure obtained from the N-best lists of the spoken questions, such that the queries for information retrieval can be better formulated. In addition, because queries formulated from the N-best results naturally generate more noisy information, we further propose to use two-layer random walk for re-ranking the retrieved webpages to produce better documents containing answers. Initial experiments performed on a set of question answering pairs in Mandarin Chinese verified that improved performance was achievable with the proposed approaches.

Shrinkage Based Features for Slot Tagging with Conditional Random Fields
Ruhi Sarikaya, Asli Celikyilmaz, Anoop Deoras, Minwoo Jeong; Microsoft, USA
Mon-P-3-5, Poster
In this paper we propose a set of class-based features that are generated in an unsupervised fashion to improve slot tagging with Conditional Random Fields (CRFs). The feature generation is based on the idea behind shrinkage based language models, where shrinking the sum of parameter magnitudes in an exponential model tends to improve performance. We use these features with CRFs and show that they consistently improve the slot tagging performance against baselines on several natural language understanding tasks. Since the proposed features are generated in an unsupervised manner without significant computational overhead, the improvements in performance comes for free and we expect that the same features may result in gains in other tagging tasks.

Cluster Based Chinese Abbreviation Modeling
Yangyang Shi, Yi-Cheng Pan, Mei-Yuh Hwang; Microsoft, China
Mon-P-3-6, Poster
Abbreviations in Chinese are widely observed in Chinese spoken language. Automatic generation of Chinese abbreviations helps to improve Chinese natural language understanding systems and Chinese search engine. The abbreviation generation is treated as a character-based tagging problem. Due to limited training data, Chinese abbreviation generation suffers from data sparseness. Two types of strategies are proposed to reduce the impact from data sparseness. First of all, in addition to using a traditional sequence labeling method — Conditional Random Fields (CRF), we propose to apply Recurrent Neural Network with Maximum Entropy Extension (RNNE) [9], which actually shows similar performance as using CRF in our experiment. Secondly, we propose to use training data clustering and latent topic modeling in abbreviation generation. Using training data clustering or topic modeling not only addresses the data sparseness, but also takes advantage of the fact that full-names from the same cluster or the same latent topic have similar abbreviation patterns. Our experimental results show that using manual clustering, the accuracy of abbreviation generation achieves relatively 8% improvement. Using Latent topics that are obtained from Latent Dirichlet Allocation (LDA), the accuracy achieves relative 10% improvement.

NOTES
Improving Named Entity Recognition with Prosodic Features
Denys Katerenchuk, Andrew Rosenberg; CUNY Graduate Center, USA
Mon-P-3-10, Poster

In natural language processing (NLP) the problem of named entity (NE) recognition in speech is well known, yet remains a challenge where performance is dependent on automatic speech recognition (ASR) system error rates. NEs are often foreign or out-of-vocabulary (OOV) words, leaving conventional ASR systems unable to recognize them. In our research, we improve a CRF-based NE recognition system by incorporating two styles of prosodic features, hypothesized ToBI labels and unsupervised clusters of acoustic features. ToBI-based features improve NE recognition by 6% absolute (F1: 0.39 v.s. F1: 0.45) on automatically recognized spontaneous speech from ACE’05.

Neural Network Models for Lexical Addressee Detection
Suman V. Ravuri¹, Andreas Stolcke²; ¹ICSI, USA; ²Microsoft, USA
Mon-P-3-11, Poster

Addressee detection for dialog systems aims to detect which utterances are directed at the system, as opposed to someone else. An important means for classification is the lexical content of the utterance, and N-gram models have been shown to be effective for this task. In this paper we investigate whether neural networks can enhance the lexical addressee detection, using data from a human-human-computer dialog system. Even though we find no improvement from simply replacing the standard N-gram LM with a neural-network LM as class likelihood estimators, improved classification accuracy can be obtained from a modified neural net model that learns distributed word representations in a first training phase, and is trained on the utterance classification task in a second phase. We obtain additional gains by combining the class likelihood estimation and classification training criteria in the second phase, and by combining multiple model architectures at the score level. Overall, we achieve over 2% absolute reduction in equal error rate over the N-gram model baseline of 27%.

Manipulating Stance and Involvement Using Collaborative Tasks: An Exploratory Comparison
Valerie Freeman, Julian Chan, Gina-Anne Levow, Richard Wright, Mari Ostendorf, Victoria Zayats; University of Washington, USA
Mon-P-3-12, Poster

The ATAROS project aims to identify acoustic signals of stance-taking in order to inform the development of automatic stance recognition in natural speech. Due to the typically low frequency of stance-taking in existing corpora that have been used to inves-
Incremental Dialog Processing in a Task-Oriented Dialog

Fabrizio Ghigi¹, Maxine Eskenazi², M. Ines Torres¹, Sunqjin Lee²; ¹Universidad del Pais Vasco, Spain; ²Carnegie Mellon University, USA

Mon-O-5-1

Incremental Dialog Processing (IDP) enables Spoken Dialogue Systems to gradually process minimal units of user speech in order to give the user an early system response. In this paper, we present an application of IDP that shows its effectiveness in a task-oriented dialog system. We have implemented an IDP strategy and deployed it for one month on a real-user system. We compared the resulting dialogs with dialogs produced over the previous month without IDP. Results show that the incremental strategy significantly improved system performance by eliminating long and often off-task utterances that generally produce poor speech recognition results. User behavior is also affected; the user tends to shorten utterances after being interrupted by the system.

Detecting Incorrectly-Segmented Utterances for Posteriori Restoration of Turn-Taking and ASR Results

Naoki Hotta¹, Kazunori Komatani¹, Satoshi Sato¹, Mikio Nakano²; ¹Nagoya University, Japan; ²Honda Research Institute Japan, Japan

Mon-O-5-2

Appropriate turn-taking is important in spoken dialogue systems as well as generating correct responses. We have developed a method that performs a posteriori restoration of incorrectly segmented utterances caused by erroneous voice activity detection (VAD), which result in automatic speech recognition (ASR) errors and inappropriate turn-taking. A crucial part of the method is to classify whether the restoration is required or not. We cast it as a binary classification problem detecting originally single utterances from pairs of utterance fragments. Various features are used representing timing, prosody, and ASR result information to improve its accuracy. Furthermore, two kinds of feature selection are performed to obtain effective and domain-independent features. The experimental results showed that the proposed method outperformed a baseline with manually-selected features by 4.8% and 3.9% in cross-domain evaluations with two domains. More detailed analysis revealed that the dominant and domain-independent features were utterance intervals and results from the Gaussian mixture model (GMM).

Segmentation and Disfluency Removal for Conversational Speech Translation

Hany Hassan, Lee Schwartz, Dilek Hakkani-Tür, Gokhan Tur; Microsoft, USA

Mon-O-5-3

In this paper we focus on the effect of on-line speech segmentation and disfluency removal methods on conversational speech translation. In a real-time conversational speech to speech translation system, on-line segmentation of speech is required to avoid latency beyond few seconds. While sentential unit segmentation and disfluency removal have been heavily studied mainly for off-line speech processing, to the best of our knowledge, the combined effect of these tasks on conversational speech translation has not been investigated. Furthermore, optimization of performance given maximum allowable system latency to enable a conversation is a newer problem for these tasks. We show that the conventional assumption of doing segmentation followed by disfluency removal is not the best practice. We propose a new approach to do simple-disfluency removal followed by segmentation and then by complex-disfluency removal. The proposed approach shows a significant gain on translation performance of up to 3 Bleu points with only 6 seconds latency to look ahead, using state-of-the-art machine translation and speech recognition systems.

Cost-Level Integration of Statistical and Rule-Based Dialog Managers

Shinji Watanabe¹, John R. Hershey¹, Tim K. Marks¹, Youichi Fujii², Yusuke Koji²; ¹MERL, USA; ²Mitsubishi Electric, Japan

Mon-O-5-4

Statistical dialog managers can potentially make more robust decisions than their rule-based counterparts, because they can account for uncertainties due to errors in speech recognition and natural language understanding. In practice, however, statistical dialog managers can be difficult to use, as they may require a large number of parameters to be inferred from limited data. Consequently, hand-crafted rule based systems are still effective for practical use. This paper proposes a method to integrate an existing rule-based dialog manager with a statistical dialog manager based on Bayes decision theory, by incorporating the rule-based dialog manager into the cost function of the statistical dialog manager. The cost function has two parts: an efficiency cost that penalizes inefficient actions, as in conventional statistical dialog approaches, and a regularization cost that slightly penalizes system actions that differ from those that would be chosen by the rule-based system. Our experiments, which use a destination-setting task in an automobile dialog scenario, demonstrate that the integrated system produces system actions that are similar to those of an existing rule-based dialog manager but enable task completion using fewer turns than the rule-based system.

Inverse Reinforcement Learning for Micro-Turn Management

Dongho Kim, Catherine Breslin, Pirros Tsiakoulis, M. Gašić, Matthew Henderson, Steve Young; University of Cambridge, UK

Mon-O-5-5

Existing spoken dialogue systems are typically not designed to provide natural interaction since they impose a strict turn-taking
Analyzing the Prosodic Characteristics of Speech-Chunks Preceding Silences in Task-Based Interactions

John Kane 1, Irena Yanushevskaya 1, Céline de Looze 1, Brian Vaughan 2, Ailbhe Ní Chasaide 1; 1Trinity College Dublin, Ireland; 2Dublin Institute of Technology, Ireland

For many applications in human-computer interaction, it is desirable to predict between-(gaps) and within-(pauses) speaker silences independently of automatic speech recognition (ASR). In this study, we focus a dataset of 6 dyadic task-based interactions and aim at automatic discrimination of gaps and pauses based on $f_0$, energy and glottal parameters derived from the speech just preceding the silence. Initial manual annotation reveals strong discriminative power of intonation tune types. In a subsequent automatic analysis using descriptive statistics of parameter contours, as well as a modelling of such contours using principal component analysis, we are able to speaker-independently predict pauses and gaps at an accuracy of 70% compared to a 56% baseline.

Oral Session 6: DNN Architectures and Robust Recognition

Peridot 202-203
14:30 – 16:30, Monday 15 September 2014
Chairs: Dong Yu and Jen-Tzung Chien

Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling

Hasim Sak, Andrew Senior, Françoise Beaufays; Google, USA

Long Short-Term Memory (LSTM) is a specific recurrent neural network (RNN) architecture that was designed to model temporal sequences and their long-range dependencies more accurately than conventional RNNs. In this paper, we explore LSTM RNN architectures for large scale acoustic modeling in speech recognition. We recently showed that LSTM RNNs are more effective than DNNs and conventional RNNs for acoustic modeling, considering moderately-sized models trained on a single machine. Here, we introduce the first distributed training of LSTM RNNs using asynchronous stochastic gradient descent optimization on a large cluster of machines. We show that a two-layer deep LSTM RNN where each LSTM layer has a linear recurrent projection layer can exceed state-of-the-art speech recognition performance. This architecture makes more effective use of model parameters than the others considered, converges quickly, and outperforms a deep feedforward neural network having an order of magnitude more parameters.

Unfolded Recurrent Neural Networks for Speech Recognition

George Saon, Hagen Soltar, Ahmad Emami, Michael Picheny; IBM T.J. Watson Research Center, USA

We introduce recurrent neural networks (RNNs) for acoustic modeling which are unfolded in time for a fixed number of time steps. The proposed models are feedforward networks with the property that the unfolded layers which correspond to the recurrent layer have time-shifted inputs and tied weight matrices. Besides the temporal depth due to unfolding, hierarchical processing depth is added by means of several non-recurrent hidden layers inserted between the unfolded layers and the output layer. The training of these models: (a) has a complexity that is comparable to deep neural networks (DNNs) with the same number of layers; (b) can be done on frame-randomized minibatches; (c) can be implemented efficiently through matrix-matrix operations on GPU architectures which makes it scalable for large tasks. Experimental results on the Switchboard 300 hours English conversational telephony task show a 5% relative improvement in word error rate over state-of-the-art DNNs trained on FMLLR features with i-vector speaker adaptation and hessian-free sequence discriminative training.

Manifold Regularized Deep Neural Networks

Vikrant Singh Tomar, Richard C. Rose; McGill University, Canada

Deep neural networks (DNNs) have been successfully applied to a variety of automatic speech recognition (ASR) tasks, both in discriminative feature extraction and hybrid acoustic modeling scenarios. The development of improved loss functions and regularization approaches have resulted in consistent reductions in ASR word error rates (WERs). This paper presents a manifold learning based regularization framework for DNN training. The associated techniques attempt to preserve the underlying low dimensional manifold based relationships amongst speech feature vectors as part of the optimization procedure for estimating network parameters. This is achieved by imposing manifold based locality preserving constraints on the outputs of the network. The techniques are presented in the context of a bottleneck DNN architecture for feature extraction in a tandem configuration. The ASR WER obtained using these networks is evaluated on a speech-in-noise task and compared to that obtained using DNN-bottleneck networks trained without manifold constraints.

Modeling Long Temporal Contexts for Robust DNN-Based Speech Recognition

Bo Li, Khe Chai Sim; National University of Singapore, Singapore

Deep Neural Networks (DNNs) have been shown to outperform traditional Gaussian Mixture Models in many Automatic Speech Recognition tasks. In this work, we investigate the potential of modeling long temporal acoustic contexts using DNNs. The complete temporal context is split into several sub-contexts. Multiple sub-context DNNs initialized with the same set of Restricted Boltzmann Machines are fine-tuned independently and their last hidden layer activations are combined to jointly predict the desired state posteriors through a single softmax output layer. From preliminary experiments on the Aurora2 multi-style training task, our proposed system models a 65-frame temporal window of speech signals and yields a 4.4% WER, outperforming the best single DNN by 12.0% relatively. With the local independence assumption,
A Long, Deep and Wide Artificial Neural Net for Robust Speech Recognition in Unknown Noise

Feipeng Li, Phani S. Nidadavolu, Hynek Hermansky; Johns Hopkins University, USA

A long deep and wide artificial neural net (LDWNN) with multiple ensemble neural nets for individual frequency subbands is proposed for robust speech recognition in unknown noise. It is assumed that the effect of arbitrary additive noise on speech recognition can be approximated by white noise (or speech-shaped noise) of similar level across multiple frequency subbands. The ensemble neural nets are trained in clean and speech-shaped noise at 20, 10, and 5 dB SNR to accommodate noise of different levels, followed by a neural net trained to select the most suitable neural net for optimum information extraction within a frequency subband. The posteriors from multiple frequency subbands are fused by another neural net to give a more reliable estimation. Experimental results show that the subband ensemble net adapts well to unknown noise.

Investigation of Deep Neural Networks for Robust Recognition of Nonlinearly Distorted Speech

Ladislav Seps, Jiri Malek, Petr Cerva, Jan Nouza; Technical University of Liberec, Czech Republic

This paper studies the use of hybrid context-dependent Deep Neural Network Hidden Markov Model (DNN-HMM) architecture for robust recognition of speech affected by real-world nonlinear distortions. We consider two types of distortions; a) signals distorted through overgained microphone preamplifier in the analog domain and b) recordings exhibiting unnatural spectral sparseness, caused by excessive denoising or low-bit-rate compression. We compare the performance of DNN-HMM architecture with that of the conventional system, based on context-dependent Gaussian Mixture Model (GMM)-HMMs, which applies channel/speaker adaptation and/or feature compensation in the front-end via Histogram Equalization (HEQ). We show that DNN-HMM architecture achieves a significantly lower Word Error Rate (WER) on the considered distorted datasets and that the obtained relative WER reduction is higher than 60%. We also investigate the usefulness of the feature compensation via HEQ for a DNN-HMM system and show that it can be helpful in the case of shallower networks.

NOTES

Oral Session 7: Speaker Recognition — Evaluation and Forensics

Peridot 204-205
14:30 - 16:30, Monday 15 September 2014
Chairs: Jean-François Bonastre and Najim Dehak

Summary and Initial Results of the 2013–2014 Speaker Recognition i-Vector Machine Learning Challenge

Désiré Bansé¹, George R. Doddington¹, Daniel Garcia-Romero², John J. Godfrey², Craig S. Greenberg¹, Alvin F. Martin¹, Alan McCree³, Mark Przybocki¹, Douglas A. Reynolds³; ¹NIST, USA; ²Johns Hopkins University, USA; ³MIT Lincoln Laboratory, USA

During late-2013 through early-2014 NIST coordinated a special i-vector challenge based on data used in previous NIST Speaker Recognition Evaluations (SREs). Unlike evaluations in the SRE series, the i-vector challenge was run entirely online and used fixed-length feature vectors projected into a low-dimensional space (i-vectors) rather than audio recordings. These changes made the challenge more readily accessible, especially to participants from outside the audio processing field. Compared to the 2012 SRE, the i-vector challenge saw an increase in the number of participants by nearly a factor of two, and a two orders of magnitude increase in the number of systems submitted for evaluation. Initial results indicate the leading system achieved an approximate 37% improvement relative to the baseline system.

Constrained Speaker Linking

David A. van Leeuwen¹, Niko Brümmer²; ¹Netherlands Forensic Institute, The Netherlands; ²AGNITiO, South Africa

In this paper we study speaker linking (a.k.a. partitioning) given constraints of the distribution of speaker identities over speech recordings. Specifically, we show that the intractable partitioning problem becomes tractable when the constraints pre-partition the data in smaller cliques with non-overlapping speakers. The surprisingly common case where speakers in telephone conversations are known, but the assignment of channels to identities is unspecified, is treated in a Bayesian way. We show that for the Dutch CGN database, where this channel assignment task is at hand, a lightweight speaker recognition system can quite effectively solve the channel assignment problem, with 93% of the cliques solved. We further show that the posterior distribution over channel assignment configurations is well calibrated.

RBM-PLDA Subsystem for the NIST i-Vector Challenge

Sergey Novoselov, Timur Pekhovsky, Konstantin Simonchik, Andrey Shulipa; Speech Technology Center, Russia

This paper presents the Speech Technology Center (STC) system submitted to NIST i-vector challenge. The system includes different subsystems based on TV-PLDA, TV-SVM, and RBM-PLDA. In this paper we focus on examining the third RBM-PLDA subsystem. Within this subsystem, we present our RBM extractor of the pseudo
i-vector. Experiments performed on the test dataset of NIST-2014 demonstrate that although the RBM-PLDA subsystem is inferior to the former two subsystems in terms of absolute minDCF, during the final fusion it provides a substantial input into the efficiency of the resulting STC system reaching 0.241 at the minDCF point.

**Limited Labels for Unlimited Data: Active Learning for Speaker Recognition**

Stephen H. Shum, Najim Dehak, James R. Glass; MIT, USA

Mon 07:45

In this paper, we attempt to quantify the amount of labeled data necessary to build a state-of-the-art speaker recognition system. We begin by using i-vectors and the cosine similarity metric to represent an unlabeled set of utterances, then obtain labels from a noiseless oracle in the form of pairwise queries. Finally, we use the resulting speaker clusters to train a PLDA scoring function, which is assessed on the 2010 NIST Speaker Recognition Evaluation. After presenting the initial results of an algorithm that sorts queries based on nearest-neighbor pairs, we develop techniques that further minimize the number of queries needed to obtain state-of-the-art performance. We show the generalizability of our methods in anecdotal fashion by applying our methods to two different distributions of utterances-per-speaker and, ultimately, find that the actual number of pairwise labels needed to obtain state-of-the-art results may be a mere fraction of the queries required to fully label the entire set of utterances.

**Bayesian Calibration for Forensic Evidence Reporting**

Niko Brümmer, Albert Swart; AGNITio, South Africa

Mon 07:55

We introduce a Bayesian solution for the problem in forensic speaker recognition, where there may be very little background material for estimating score calibration parameters. We work within the Bayesian paradigm of evidence reporting and develop a principled probabilistic treatment of the problem, which results in a Bayesian likelihood-ratio as the vehicle for reporting weight of evidence. We show in contrast, that reporting a likelihood-ratio distribution does not solve this problem. Our solution is experimentally exercised on a simulated forensic scenario, using NIST SRE'12 scores, which demonstrates a clear advantage for the proposed method compared to the traditional plugin calibration recipe.

**Replicate Mismatch Between Test/Background and Development Databases: The Impact on the Performance of Likelihood Ratio-Based Forensic Voice Comparison**

Shunichi Ishihara; Australian National University, Australia

Mon 08:05

In this paper, we report on a study which demonstrates that the mismatch in within-speaker replicate numbers (the number of tokens used to model each sample) between test/background and development databases has a large impact on the performance of a forensic voice comparison (FVC) system. We describe how and to what extent the different degrees of the mismatch influence the performance of the FVC system. The performance of an FVC system based on temporal MFCC features and the Multivariate Kernel Density Likelihood Ratio procedure is tested in terms of its validity and reliability under the mismatched conditions.

**Oral Session 8: Speech Production I**

Peridot 201

14:30 – 16:30, Monday 15 September 2014

Chairs: Pascal Perrier and Takayuki Arai

**Automatic Estimation of the Lip Radiation Effect in Glottal Inverse Filtering**

Manu Aittrakshinen¹, Tom Bäckström², Paavo Alku¹; ¹Aalto University, Finland; ²FAU Erlangen-Nürnberg, Germany

Mon 08:15

In the analysis of speech production, glottal inverse filtering has proved to be an effective yet non-invasive method for obtaining information about the voice source. One of the main challenges of the existing methods is blind estimation of the contribution of the lip radiation, which must often be manually determined. To obtain a fully automatic system, we propose an automatic method for determining the lip radiation parameter. Our method is based on a physically-motivated quality criteria for the glottal flow, which can be approximated by minimization of the norm-1. Experiments show that the parameters obtained by the automatic method are mostly within the 95% confidence intervals of the mean values obtained by manual tuning by experts.

**Simulation of 3D Larynges with Asymmetric Distribution of Viscoelastic Properties in their Vocal Folds**

Marcelo de Oliveira Rosa; UTFPR, Brazil

Mon 08:24

Here 3D larynx models whose viscoelastic properties of the superficial tissues of the true vocal folds were altered in specific regions at the glottis — producing different left-right fold asymmetries — are analyzed. The idea is to initiate studies about the dynamic of larynges under possible pathological conditions. All simulations used a strategy based on the finite element method to solve a fluid-structure interaction (airflow in the larynx) with a contact-impact problem (glottal closure). The results show that glottal closure is directly influenced by the presence and location of a stiffer mass in the surface of one of the vocal folds. Changes in glottal signal $F_0$ (up to +13.7Hz), open quotient (up to 0.5812), and relative jitter (up to 15.2381%) were observed and are discussed here.

**Comparison of Vocal Tract Transfer Functions Calculated Using One-Dimensional and Three-Dimensional Acoustic Simulation Methods**

Hironori Takemoto¹, Parham Mokhtari¹, Tatsuya Kitamura²; ¹NICT, Japan; ²Kanoun University, Japan

Mon 08:33

Acoustic characteristics of the vocal tract have been investigated extensively in the literature using a one-dimensional (1D) acoustic simulation method. Because the 1D method assumes plane wave propagation only, it is recognized to be valid only in the low frequency region (below about 4 or 5 kHz). Recently, a three-dimensional (3D) acoustic simulation method was developed, to obtain more precise acoustic characteristics of the vocal tract. In the present study, from a male’s vocal tract shapes, transfer

**Notes**
functions were calculated using the 1D and 3D methods and compared with each other to evaluate the valid frequency range of the 1D method. As a result, when acoustic effects of the piriform fossae were considered in the 1D method, the transfer functions disagreed with each other up to 7 kHz (ignoring small dips). The 3D method showed that a deep dip was generated at around 8 kHz by the transverse resonance mode in the pharynx. Above this dip frequency, the transfer functions disagreed with each other. Thus, the 1D method is valid up to 7 kHz for this subject. Because this subject has a relatively large vocal tract, in general the upper limit of the valid frequency range could exceed 8 kHz.

A Study of Invariant Properties and Variation Patterns in the Converter/Distributor Model for Emotional Speech

Jangwon Kim, Donna Erickson, Sungbok Lee

Invertible properties of vocal organ controls at an abstract level are crucial for better understanding and modeling of the speech production mechanism. Despite the large variability of articulatory movements at the execution level, the Converter/Distributor (C/D) model provides a systematic and comprehensive framework for the prosodic organization of speech production, based on the invariant properties of articulatory movements with the concept of "iceberg" region. The goal of this paper is two-fold: (i) to examine the invariant properties in the C/D model in emotional speech, and (ii) to understand emotion-dependent variation patterns of important parameters in the C/D model framework. Experimental results support the validity of strong linear relationship between the speed and excursion of critical articulators at the iceberg points for emotional speech. Also, emotion-dependent variation patterns of the C/D model parameters, (e.g., relatively smaller "shadow" angle and greater syllable magnitude for happiness) are reported. Finally, the emotion-dependent relationships between the abstract-level C/D model parameters and the surface-level parameters of the invariant articulatory behaviors are reported.

A Hybrid Approach to 3D Tongue Modeling from Vocal Tract MRI Using Unsupervised Image Segmentation and Mesh Deformation

Alexander Hewer, Ingmar Steiner, Stefanie Wuhrer

Vocal tract magnetic resonance imaging (MRI) has become one of the preferred imaging modalities for the analysis of human speech production. However, the raw image data must be segmented before further analysis can take place. This paper describes a hybrid approach to extract a 3D tongue model from 3D or 2D MRI scans of the vocal tract during speech, which combines unsupervised image segmentation with a mesh deformation technique. An efficient, minimally supervised segmentation algorithm can also be used as an alternative to provide a robust fallback in certain isolated cases. Both image segmentation algorithms produce a point cloud, which is completed and registered by deforming a template mesh to the data. Since the mesh deformation can be applied even with a sparse point cloud, it is possible to extract realistic 3D tongue shapes even from the 2D video frames of real-time MRI. Our approach is applied to several sets of available MRI data and yields promising results.

Notes

Estimation of Vocal-Tract Shape from Speech Spectrum and Speech Resynthesis Based on a Generative Model

Tokihiro Kaburagi - Kyushu University, Japan

Precise control of articulatory parameters is difficult and prevents a physical model from generating natural sounding speech signals. To determine vocal-tract shape from speech, this paper presents an inversion method for simultaneously estimating the cross-sectional area and length of the vocal tract. In addition, we performed speech resynthesis from a time-series of estimated vocal-tract shapes. The vocal-tract shape is determined through an iterative procedure that gradually optimizes the parameter values to produce the target speech spectrum. The vocal-tract shape is updated using a sensitivity function that represents the change in formant frequency caused by a small perturbation of the vocal-tract shape. When combined with a perturbation relationship of speech spectrum parameters (i.e., cepstrum parameters) and formants, our method effectively optimizes the vocal-tract shape. We quantitatively examined the accuracy using area function data for 10 isolated vowels. The results showed that the average area error was 0.43 cm² and the average length error was 0.23 cm. This indicates that the vocal-tract shape was determined with satisfactory accuracy. We also performed an estimation experiment for continuous speech and synthesized speech from the estimated vocal-tract shape.

Special Session 2(a): INTERSPEECH 2014 Computational Paralinguistics Challenge (ComParE) I

The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load

Björn Schuller, Stefan Steidl, Anton Batliner, Julien Epps, Florian Eyben, Fabien Ringeval, Erik Marchi, Yue Zhang; Imperial College London, UK; FAU Erlangen-Nürnberg, Germany; Technical University Munich, Germany; University of New South Wales, Australia

The INTERSPEECH 2014 Computational Paralinguistics Challenge provides for the first time a unified test-bed for the automatic recognition of speakers' cognitive and physical load in speech. In this paper, we describe these two Sub-Challenges, their conditions, baseline results and experimental procedures, as well as the COMPARE baseline features generated with the openSMILE toolkit and provided to the participants in the Challenge.

Filtering and Subspace Selection for Spectral Features in Detecting Speech Under Physical Stress

Jouni Pohjalainen, Paavo Alku; Aalto University, Finland

This paper investigates approaches to modeling the time evolution of short-time spectral features in paralinguistic speech type classification, where we focus on detection of speech influenced by physical exertion. The time series model consists of autoregressive...
processes of multiple time scales and orders and is trained to describe the long-term dynamics of a given target speech class. The model is applied in two ways in improving long-term modeling in the detection task: 1) to perform predictive filtering of the features and 2) to automatically select instantaneous classification subspaces. The spectrum analysis method underlying the short-time features is also varied between the standard discrete Fourier transform and a time-weighted linear predictive method which yields smooth all-pole spectrum envelope models. Configurations of the proposed methods are evaluated in the Physical Load task of the Interspeech 2014 Computational Paralinguistics Challenge and show improvement over the baseline timbral classifier and the challenge baseline. Also the interrelationships among the methods are discussed.

Automatic Recognition of Speaker Physical Load Using Posterior Probability Based Features from Acoustic and Phonetic Tokens
Ming Li; Sun Yat-sen University, China
Mon-SPl-2a-3

This paper presents an automatic speaker physical load recognition approach using posterior probability based features from acoustic and phonetic tokens. In this method, the tokens for calculating the posterior probability or zero-order statistics are extended from the conventional MFCC trained Gaussian Mixture Models (GMM) components to parallel phonetic phonemes and tandem feature trained GMM components. Phoneme recognizers from five different languages are employed to extract the phoneme posterior probabilities. We show that these histogram style features at both the acoustic and phonetic levels are effective and complementary for capturing the speaker physical load information from short utterances. Support vector machine is adopted as the supervised classifier. By combining the proposed methods with the OpenS-MILE baseline which covers the acoustic and prosodic information further improves the final performance. The proposed fusion system achieves 70.18% and 72.81% unweighted accuracy on the validation and test set of the Munich Bio-voice Corpus for the binary physical load level recognition task in the INTERSPEECH 2014 Computational Paralinguistics Challenge.

Canonical Correlation Analysis and Local Fisher Discriminant Analysis Based Multi-View Acoustic Feature Reduction for Physical Load Prediction
Haysem Kaya, Tuğçe Özkaplan, Albert Ali Salah, Sadik Fikret Gürgen; Boğaziçi Üniversitesi, Turkey
Mon-SPl-2a-4

In this study we present our system for INTERSPEECH 2014 Computational Paralinguistics Challenge (ComParE 2014), Physical Load Sub-challenge (PLS). Our contribution is twofold. First, we propose using Low Level Descriptor (LLD) information as hints, so as to partition the feature space into meaningful subsets called views. We also show the virtue of commonly employed feature projections, such as Canonical Correlation Analysis (CCA) and Local Fisher Discriminant Analysis (LFDA) as ranking feature selectors. Results indicate the superiority of multi-view feature reduction approach to its single-view counterpart. Moreover, the discriminative projection matrices are observed to provide valuable information for feature selection, which generalize better than the projection itself. In our preliminary experiments we reached 75.35% Unweighted Average Recall (UAR) on PLS test set, using CCA based multi-view feature selection.

Ensemble of Machine Learning Algorithms for Cognitive and Physical Speaker Load Detection
How Jing 1, Ting-Yao Hu 2, Hung-Shin Lee 3, Wei-Chen Chen 3, Chi-Chun Lee 3, Yu Tsao 1, Hsin-Min Wang 1; 1Academia Sinica, Taiwan; 2National Taiwan University, Taiwan; 3National Tsing Hua University, Taiwan
Mon-SPl-2a-6

We present our methods and results on participating in the Interspeech 2014 Computational Paralinguistics Challenge (ComParE) of which the goal is to detect certain type of load of a speaker using acoustic features. There are in total seven classification models contributing to our final prediction, namely, neural network with rectified linear unit and dropout (ReLU-Net), conditional restricted Boltzmann machine (CRBM), logistic regression (LR), support vector machine (SVM), Gaussian discriminant analysis (GDA), k-nearest neighbors (KNN), and random forest (RF). When linearly blending the predictions of these models, we are able to get significant improvements over the challenge baseline.

Detecting the Intensity of Cognitive and Physical Load Using AdaBoost and Deep Rectifier Neural Networks
Gábor Gosztolya, Tamás Grósz, Róbert Busa-Fekete, László Tóth; MTA-SZTE RGAI, Hungary
Mon-SPl-2a-7

The Interspeech ComParE 2014 Challenge consists of two machine learning tasks, which have quite a small number of examples. Due to our good results in ComParE 2013, we considered AdaBoost a suitable machine learning meta-algorithm for these tasks, besides we also experimented with Deep Rectifier Neural Networks. These differ from traditional neural networks in that the former have several hidden layers, and use rectifier neurons as hidden units. With AdaBoost we achieved competitive results, whereas with the neural networks we were able to outperform baseline SVM scores in both Sub-Challenges.

Poster Session 4: Hearing and Perception
Max Atria Gallery
14:30 – 16:30, Monday 15 September 2014
Chair: Tan Lee

Revisiting the Right-Ear Advantage for Speech: Implications for Speech Displays
Nandini Iyer 1, Eric Thompson 2, Brian Simpson 1, Griffin Romigh 1; 1AFRL, USA; 2Ball Aerospace, USA
Mon-P-4-1, Poster

Cherry (1953) reported that when listeners were presented with dichotic signal over headphones, they could reliably report words presented to the attended ear, while only being aware of the gross properties of the talker in the unattended ear. More recently, Gallun et al. (2007) showed that there were large differences in performance on dichotic tasks depending on ear of presentation, with significantly larger errors occurring when the target was presented to the left, rather than right ear (i.e., a right-ear advantage). In the current experiment, we explored two factors, type of signal in the non-target ear and uncertainty about the target ear, and their effects on right-ear advantage. The results indicated that the right-ear advantage was modulated by two factors: 1) nature of the speech stimuli presented in the unattended ear, and 2) target
Comparing Reaction Time Sequences from Human Participants and Computational Models
L. ten Bosch, M. Ernestus, L. Boves; Radboud Universiteit Nijmegen, The Netherlands
Mon-P-4-2, Poster
This paper addresses the question how to compare reaction times computed by a computational model of speech comprehension with observed reaction times by participants. The question is based on the observation that reaction time sequences substantially differ per participant, which raises the issue of how exactly the model is to be assessed. Part of the variation in reaction time sequences is caused by the so-called local speed: the current reaction time correlates to some extent with a number of previous reaction times, due to slowly varying variations in attention, fatigue etc. This paper proposes a method, based on time series analysis, to filter the observed reaction times in order to separate the local speed effects. Results show that after such filtering the between-participant correlations increase as well as the average correlation between participant and model increases. The presented technique provides insights into relevant aspects that are to be taken into account when comparing reaction time sequences.

Detecting the Number of Competing Speakers — Human Selective Hearing versus Spectrogram Distance Based Estimator
Valentin Andrei, Horia Cucu, Andi Buzo, Corneliu Burileanu; Universitatea Politehnica din București, Romania
Mon-P-4-3, Poster
This study describes an experiment designed to establish the maximum number of competing speakers that can be detected accurately by a human listener and compares the results with the ones produced by using a distance based estimator working in frequency domain. We mixed a set of high quality audio samples with continuous speech, produced by publicly known people (actors, journalists and politicians) and also unknown persons and then we played the tracks to each listener within a target group. The volunteers were asked how many cumulated speakers they counted and how they obtained the response. We observed that while human subjects showed a correct detection ratio of 31%, we were able to establish a set of equally spaced thresholds for the estimator in order to achieve 66% accuracy. The paper also summarizes the methods that were reported by the listeners to help in the detection.

The Influence of Sensory Memory and Attention on the Context Effect in Talker Normalization
Guo Li, Gang Peng; Chinese University of Hong Kong, China
Mon-P-4-4, Poster
Studies on talker normalization have reported that contexts would affect how a vowel/consonant/tone is perceived, which suggests that listeners use cues in the context as a reference for speech perception. However, it is unclear how the cues of the context are encoded in memory, and to what extent the context effects are influenced by interruption in sensory memory and reduction in attentional resources. To fill in this gap, this study examined the effects of noise interruption and a secondary visual task on the identification of words carrying level tones in context by native Cantonese speakers. Experiment 1 compared the tone identification performance between a block where a 300ms noise was presented immediately after the context to interrupt the normalization effects of the context on the following target word and a block without noise interruption. Experiment 2 compared the tone identification performance between a block where participants had to perform a secondary visual task (picture same/different discrimination) and a block without a secondary visual task. Results suggest that context cues are likely encoded both in sensory memory and in short-term categorical memory and that reduction in attentional resources has marginal influence on tone normalization.

Automatic Speech Recognition with Primarily Temporal Envelope Information
Payton Lin 1, Fei Chen 2, Syu Siang Wang 1, Ying-Hui Lai 1, Yu Tsao 1; 1Academia Sinica, Taiwan; 2University of Hong Kong, China
Mon-P-4-5, Poster
The aim of this study is to devise a computational method to predict cochlear implant (CI) speech recognition. Here, we describe a high-throughput screening system for optimizing CI speech processing strategies using hidden Markov model (HMM)-based automatic speech recognition (ASR). Word accuracy was computed on vocoded CI speech synthesized from primarily multi-channel temporal envelope information. The ASR performance increased with the number of channels in a similar manner displayed in human recognition scores. Results showed the computational method of HMM-based ASR offers better process control for comparing signal carrier type. Training-test mismatch reduction provided a novel platform for reevaluating the relative contributions of spectral and temporal cues to human speech recognition.

An Adaptive Envelope Compression Strategy for Speech Processing in Cochlear Implants
Ying-Hui Lai 1, Fei Chen 2, Yu Tsao 1; 1Academia Sinica, Taiwan; 2University of Hong Kong, China
Mon-P-4-6, Poster
Hearing-impaired patients have limited hearing dynamic range for speech perception, which partially accounts for their poor speech understanding abilities, particularly in noise. Wide dynamic range compression aims to compress speech signal into the usable hearing dynamic range of hearing-impaired listeners; however, it normally uses a static compression based strategy. This work proposed a strategy to continuously adjust the envelope compression ratio for speech processing in cochlear implants, named adaptive envelope compression (AEC) strategy. This AEC strategy aims to keep the compression processing as close to linear as possible, while still confine the compressed amplitude envelope within the pre-set dynamic range. Vocoder simulation experiments showed that, when narrowed down to a small dynamic range, the intelligibility of AEC-processed sentences was significantly better than those processed by static envelope compression. This makes the proposed AEC strategy a promising way to improve speech recognition performance for implanted patients in the future.

Notes
Articulatory Dynamics and Coordination in Classifying Cognitive Change with Preclinical mTBI

Brian S. Helfer 1, Thomas F. Quatieri 1, James R. Williamson 1, Laurel Keyes 1, Benjamin Evans 1, W. Nicholas Greene 1, Trina Vian 1, Joseph Lacigronola 1, Trey Shenk 2, Thomas Talavage 2, Jeff Palmer 1, Kristin Heaton 3; 1MIT Lincoln Laboratory, USA; 2University of Colorado at Boulder, USA; 3USARIEM, USA

Investigation of the Relative Perceptual Importance of Temporal Envelope and Temporal Fine Structure Between Tonal and Non-Tonal Languages

Dongmei Wang 1, James M. Kates 2, John H.L. Hansen 1; 1University of Texas at Dallas, USA; 2University of Colorado at Boulder, USA

Speech analysis has shown potential for identifying neurological impairment. With brain trauma, changes in brain structure or connectivity may result in changes in source, prosodic, or articulatory aspects of voice. In this work, we examine the articulatory components of speech reflected in formant tracks, and how changes in track dynamics and coordination map to cognitive decline. We address a population of athletes regularly receiving impacts to the head and showing signs of preclinical mild traumatic brain injury (mTBI), a state indicated by impaired cognitive performance occurring prior to concussion. We hypothesize that this preclinical damage results in 1) changes in average vocal tract dynamics measured by formant frequencies, their velocities, and acceleration, and 2) changes in articulatory coordination measured by a novel formant-frequency cross-correlation characterization. These features allow machine learning algorithms to detect preclinical mTBI identified by a battery of cognitive tests. A comparison is performed of the effectiveness of vocal tract dynamics features versus articulatory coordination features. This evaluation is done using receiver operating characteristic (ROC) curves along with confidence bounds. The articulatory dynamics features achieve area under the ROC curve (AUC) values between 0.72 and 0.98, whereas the articulatory coordination features achieve AUC values between 0.94 and 0.97.

A Hearing Impairment Simulation Method Using Audiogram-Based Approximation of Auditory Characteristics

Nozomi Jinbo, Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakrianti Sakti, Satoshi Nakamura; NAIST, Japan

In this paper, we investigate the relative perceptual importance of the temporal envelop (TE) and temporal fine structure (TFS) between tonal language and non-tonal language perception. The “auditory chimera” experiment is conducted on both American English and Mandarin Chinese with the same conditions. Our experimental results showed that there is no significant difference between Mandarin Chinese and American English when the interference noise is speech-shaped noise. However, there is a distinct relative perceptual importance difference when the interference noise is a competing speech. The results also show that speech perception in Mandarin Chinese language is more sensitive to TFS distortion than is American English. For addition, with fewer auditory channels (one or two bands), speech perception in American English is more sensitive to TE distortion compared to Mandarin Chinese. Finally, the related discussions are made according to the perceptual difference between tonal language and non-tonal language regarding the speech signal processing strategies.

Vowel Spectral Contributions to English and Mandarin Sentence Intelligibility

Daniel Fogerty 1, Fei Chen 2; 1University of South Carolina, USA; 2University of Hong Kong, China

Hearing impairment simulation is an effective technique to educate normal-hearing people about auditory perception of the hearing-impaired. Because auditory characteristics of the hearing impaired vary greatly from person-to-person, personalization of the hearing impairment simulation systems is essential to accurately simulate these individual differences. However, measurement of auditory characteristics of individuals is time-consuming work. In this paper, we propose a hearing impairment simulation method that is easily applied to individual hearing-impaired persons. Auditory filter characteristics and gain characteristics are estimated from easily measurable audiograms of each individual. We also implement a method for manually adjusting the hearing impairment level to improve accuracy of the proposed hearing impairment simulation. An experimental evaluation is conducted to compare intelligibility between hearing-impaired and normal-hearing persons with the proposed hearing impairment simulation. The experimental results show that the proposed method effectively makes the word correct rate and phoneme confusion tendency of the normal hearing persons similar to those of the hearing impaired persons.

Significance of Aperiodicity in the Pitch Perception of Expressive Voices

Vinay Kumar Mittal, B. Yegnanarayana; IIT Hyderabad, India

In this paper, we study the significance of aperiodicity in the pitch perception of expressive voices such as Noh voice and laughter signals. The excitation source characteristics in the production of these signals is represented in terms of a sequence of impulses. The impulse sequence is derived from the acoustic signal using a modified zero-frequency filtering method. The time intervals between successive impulses and relative amplitudes of impulses are related to the presence of subharmonics and pitch-perception in expressive voices. The role of aperiodicity and subharmonics
in the perception of distinct voice quality of expressive voices is examined. The significance of aperiodicity is also analysed by synthesis, using two synthetic AM/FM sequences for excitation. Saliency is used as a measure of pitch perception. The $F_0$ extraction using this pitch perception information for expressive voices is also demonstrated.

**Poster Session 5: Cross-Linguistic Studies**

Max Atria Gallery
14:30 – 16:30, Monday 15 September 2014
Chair: Odette Scharenborg

**DIAPIX-FL: A Symmetric Corpus of Problem-Solving Dialogues in First and Second Languages**

Mirjam Wester$^1$, María Luisa García Lecumberri$^2$, Martin Cooke$^3$; $^1$University of Edinburgh, UK; $^2$Universidad del País Vasco, Spain; $^3$Ikerbasque, Spain

This paper describes a corpus of conversations recorded using an extension of the DiapixUK task: the Diapix Foreign Language corpus (DIAPIX-FL). English and Spanish native talkers were recorded simultaneously in both English and Spanish. The bidirectionality of the corpus makes it possible to separate language (English or Spanish) from speaking in a first language (L1) or second language (L2). An acoustic analysis was carried out to analyse changes in $F_0$, voicing, intensity, spectral tilt and formants that might result from speaking in an L2. The effect of L1 and nativeness on turn types was also studied. Every word was segmented, including pauses, elisions, and incomplete words. Speakers displayed certain patterns that suggest an on-going process of L2 phonological acquisition, such as the overall percentage of voicing in their speech. Results also show an increase in hesitation phenomena (pauses, elisions, incomplete turns), a decrease in produced speech and speech rate, a reduction of $F_0$ range, raising of minimum $F_0$ when speaking in the non-native language which are consistent with more tentative speech and may be used as indicators of non-nativeness.

**Cross-Linguistic Investigations of Oral and Silent Reading**

Christophe Coupé, Yoon Mi Oh, François Pellegrino, Egidio Marsico; DDL (UMR 5596), France

Recent research on speech rate (Pellegrino et al., 2011) has shown that languages differ in terms of syllable rate, and that these differences are compensated by the average amount of information carried by syllables. The more syllables a language needs to express a given amount of information, the higher its syllable rate tends to be.

These results were obtained with subjects reading texts on a computer screen. The question arose whether silent reading rates would correlate with oral reading rates across languages. Although silent and oral reading fluency has been studied with respect to reading comprehension and how it develops in children, little literature focuses on comparing them in different languages.

We present here data for 8 languages (Cantonese, Finnish, French, Japanese, Korean, Mandarin, Serbian, and Thai). For each language, silent and oral syllables rates as well as reading durations were measured for several subjects and 15 different texts. Various comparisons were performed, and mixed-effects regression models were used to further evaluate the weight of the different variables (gender, language, speaker and text).

**Non-Native Word Recognition in Noise: The Role of Word-Initial and Word-Final Information**

Juul Coumans, Roeland van Hout, Odette Scharenborg; Radboud Universiteit Nijmegen, The Netherlands

When listening in noisy conditions, word recognition seems to be much harder in a non-native language than in one’s native language. Native listeners use both word-initial and word-final information for word recognition in clean listening conditions, whereas word-initial information is the most important. When listening in noise, however, word-final information becomes relatively more important. This study investigates whether non-native listeners are able to use word-initial and word-final information when recognizing words in noise, and whether these information sources are equally important when listening conditions become increasingly harder. Forty-seven Dutch students participated in an English word recognition experiment, where either a word’s onset or offset was masked by speech-shaped noise with different signal-to-noise ratios. The results showed that non-native listeners are able to use both word-initial and word-final information for word recognition, but fewer words were recognized with increasing difficulty of the listening conditions when the onset of words was masked. Thus, word-initial information is more important than word-final information for word recognition when listening conditions become harder. This increasing effect occurred independently from the proficiency level in the non-native language of the participants, although proficiency level was correlated to test performance in general.

**The Effects of High and Low Variability Phonetic Training on the Perception and Production of English Vowels /e/-/æ/ by Cantonese ESL Learners with High and Low L2 Proficiency Levels**

Janice Wing Sze Wong; Hong Kong Baptist University, China

This study investigated the effects of two perceptually-based training paradigms in both the perception and production of English /e/ and /æ/ by Cantonese ESL learners with high and low listening and oral proficiency levels. Sixty-four subjects participated in the study, in which 22 (9 with high proficiency, H-HV; 13 with low proficiency, L-HV) were trained under High Variability Phonetic Training (HVPT) approach, 19 (8 with high proficiency, H-LV; 11 with low proficiency, L-LV) were trained under Low Variability Phonetic Training (LVPT) approach whereas 23 (10 with high proficiency, H-CO; 13 with low proficiency, L-CO) were the control subjects. Both training approaches were effective in improving the subjects’ perception of the two vowels, with HVPT groups showing more robust improvement than LVPT groups. Perceptual learning could also be generalized to new words and new speakers and be transferred to the production domain, with HVPT groups outperforming LVPT groups. However, subjects with different proficiency levels learned to similar degrees in all tests. The results demonstrated that both approaches offered a type of learning that allows attention to focus on phonetic information, which is different from what is learned in an L2 classroom; whereas stimulus variability also plays a role in the learning.

NOTES
In this paper we present a study on Dutch vowel production by Spanish learners that was carried out within the framework of our research on Computer Assisted Pronunciation Training (CAPT). The aim of this study was to obtain detailed information on production of Dutch vowels by Spanish learners, which can be employed to develop effective CAPT programs for this specific target group. We collected speech from learners with varying proficiency levels (A1 - B2 of the CEFR), which was transcribed, segmented and acoustically analyzed. We present data on the frequency of pronunciation errors and on detailed analyses of duration and acoustic properties of the vocalic realizations. The results indicate that Spanish learners of Dutch have difficulties in realizing several Dutch vowel contrasts and that they differ from native speakers in the way they employ duration and spectral properties to realize these contrasts. We discuss these results in relation to those of previous studies on Dutch vowel perception by Spanish listeners and relate them to current theories on speech learning.

English Consonant Confusions by Greek Listeners in Quiet and Noise and the Role of Phonological Short-Term Memory

Angelos Lengeris, Katerina Nicolaidis; AUTH, Greece

This study investigated English consonant identification by Greek listeners and the role of phonological short-term memory (PSTM) in listeners’ identification ability. Twenty Greek university students who had received formal instruction in English identified 24 English consonants (embedded in VCV syllables) presented in quiet and in two noise types, a competing talker at a signal-to-noise ratio (SNR) of -6dB and an 8-speaker babble at an SNR of -2dB. Participants’ PSTM was assessed via a serial non-word recognition task in Greek. The results showed that identification scores in quiet were significantly higher than in noise. There was no difference in scores between the two noise conditions. PSTM correlated with English consonant identification in quiet and in the two types of noise; listeners with greater PSTM capacity were also better in identifying English consonants in quiet and noise, a finding that extends previous research in quiet to L2 perception in adverse listening conditions. English consonant confusion patterns are interpreted as caused by a combination of first-language interference (at both the phonetic and phonological levels) and spectral/articulatory factors.

Corpus-Based L2 Phonological Data and Semi-Automatic Perceptual Analysis: The Case of Nasal Vowels Produced by Beginner Japanese Learners of French

Sylvain Detey¹, Isabelle Racine², Julien Eychenne³, Yuji Kawaguchi⁴; ¹Waseda University, Japan; ²Université de Genève, Switzerland; ³Hankuk University of Foreign Studies, Korea; ⁴Tokyo University of Foreign Studies, Japan

Processing large amounts of non-native (L2) phonological data for acquisition-related research remains a challenging task, especially when acoustic analyses are not straightforward as is the case with nasal vowels. Within the InterPhonologie du Français Contemporain project (IPFC), we have developed a perceptual coding procedure and a piece of dedicated software aimed at providing an intermediate stage between fine-grained acoustic analyses and coarse-grained phonological categorization, such as 'substitution' or 'deletion', of non-native productions. Our code allows us to examine the left and right phonological contexts of the segment under scrutiny and assess the nasality, quality and potential consonantal extracencies of the non-native nasal vowels. We have applied this procedure to Japanese data collected in a longitudinal study of French interphonology, focusing on the vowels /A/-, /O/-, /E/- produced by 22 beginner university students in a wordlist repetition and reading task. Our study reveals an overall good production rate in terms of nasality for such beginner learners but also a lower rate of quality accuracy for the three vowels, as well as better performances in the repetition task. We discuss our results in light of current L2 learning theories and the phonetic-phonological characteristics of Japanese.

Perception of Prosodic Prominence and Boundaries by L1 and L2 Speakers of English

Gábor Pintér¹, Shinobu Mizuguchi¹, Koichi Tateishi²; ¹Kobe University, Japan; ²Kobe College, Japan

This study provides a direct comparison of boundary and prominence perception strategies between Japanese EFL learners and native speakers of English using the Rapid Prosody Transcription (RPT) method. Although RPT experiments are available for both native English speakers [1], [2], [3] and Japanese EFL learners [4], a direct comparison of the available data is problematic as the stimuli sets used in the experiments are not identical. The present research addresses this issue by using identical stimuli sets across L1 and L2 listeners. The data for native English speakers was taken from RPT experiments carried out by Jennifer Cole with Yoonsook Mo and colleagues [1-3, 5-8]. The non-native data was collected by re-running a subset of RPT tasks reported in the work by Cole and Mo with 108 Japanese undergraduate students. Although native speakers perceived more boundaries and prominent words than L2 speakers, the results outlined surprisingly similar perceptual strategies. A strong correlation was found between the responses of native speakers and Japanese learners of English in both boundary and prominence perception tasks. In boundary perception both groups relied heavily on silent pauses and vocal fillers. In prominence detection the responses correlated with vowel duration, maximum amplitude, and maximum pitch in this specific order for both language groups.

Prosody Perception, Reading Accuracy, Nonliteral Language Comprehension, and Music and Tonal Pitch Discrimination in School Aged Children

Rose Thomas Kalathottukaren, Suzanne C. Purdy, Elaine Ballard; University of Auckland, New Zealand

Twenty-five school aged children with normal hearing were tested on their perception of prosody using the receptive prosody subtests of the Profiling Elements of Prosody in Speech-Communication (PEPS-C) and Child Paralanguage subtest of Diagnostic Analysis of Non Verbal Accuracy 2 (DANVA 2). Performance of four children with hearing loss on the two prosody measures was compared with performance of normal hearing children. Children were also tested on their reading accuracy, comprehension of nonliteral language, and music and tonal pitch discrimination. Overall

NOTES
results showed that younger children aged 7;1 to 9;11 years had significantly poorer scores than 10;1 to 12;11 year olds on the Contrastive Stress Recognition subtest of PEPs-C and the DANVA 2. Child Paralanguage subtest, indicating a developmental effect on speech prosody perception. Children with hearing loss had poorer scores and greater variability on PEPs-C and DANVA 2 assessments compared to normal hearing controls. Statistically significant correlations were observed between prosody perception scores and musical pitch perception and reading measures for the normal hearing group. This is consistent with previous studies showing links between reading and prosody perception [7,8]. Significant correlation between prosody perception and musical pitch discrimination indicates that pitch is an important cue for prosody perception.

**Phoneme Category Retuning in a Non-Native Language**

Polina Drozdova, Roeland van Hout, Odette Scharenborg; Radboud Universiteit Nijmegen, The Netherlands

Previous studies have demonstrated that native listeners modify their interpretation of a speech sound when a talker produces an ambiguous sound in order to quickly tune into a speaker, but there is hardly any evidence that non-native listeners employ a similar mechanism when encountering ambiguous pronunciations. So far, one study demonstrated this lexically-guided perceptual learning effect for nonnatives, using phoneme categories similar in the native language of the listeners and the non-native language of the stimulus materials. The present study investigates the question whether phoneme category retuning is possible in a nonnative language for a contrast, /r/-/r/, which is phonetically differently embedded in the native (Dutch) and nonnative (English) languages involved. Listening experiments indeed showed a lexically-guided perceptual learning effect. Assuming that Dutch listeners have different phoneme categories for the native Dutch and non-native English /r/, as marked differences between the languages exist for /r/, these results, for the first time, seem to suggest that listeners are not only able to retune their native phoneme categories but also their non-native phoneme categories to include ambiguous pronunciations.

**Speech Emotion Recognition with Cross-Lingual Databases**

Bo-Chang Chiu, Chia-Ping Chen; National Sun Yat-sen University, Taiwan

In this paper, we investigate cross-lingual automatic speech emotion recognition. The basic idea is that the emotion recognition system is based on the acoustic features only, it is possible to combine data in different languages to improve the recognition accuracy. We begin with the construction of a Mandarin database of emotional speech, which is similar to the well-known Berlin Database of Emotional Speech (EMO-DB) in the composition and size. In order to reduce the variability due to different languages and different speakers, we propose to apply histogram equalization as a data normalization method. Recognition systems based on support vector machines have been evaluated on EMO-DB. Compared to the baseline system without multi-lingual databases and data normalization, the proposed system has achieved a relative improvement of 39.9% in the emotion recognition accuracy, from 86.2% to 91.7%. The accuracy is among the best known results reported on EMO-DB, if not the best.

---

**Poster Session 6: Speaker Diarization**

Max Atria Gallery  
14:30 - 16:30, Monday 15 September 2014  
Chair: Corinne Fredouille

**Speaker Diarization Using Eye-gaze Information in Multi-Party Conversations**

Koji Inoue, Yukoh Wakabayashi, Hiromasa Yoshimoto, Tatsuya Kawahara; Kyoto University, Japan

We present a novel speaker diarization method by using eye-gaze information in multi-party conversations. In real environments, speaker diarization or speech activity detection of each participant of the conversation is challenging because of distant talking and ambient noise. In contrast, eye-gaze information is robust against acoustic degradation, and it is presumed that eye-gaze behavior plays an important role in turn-taking and thus in predicting utterances. The proposed method stochastically integrates eye-gaze information with acoustic information for speaker diarization. Specifically, three models are investigated for multi-modal integration in this paper. Experimental evaluations in real poster sessions demonstrate that the proposed method improves accuracy of speaker diarization from the baseline acoustic method.

**Unsupervised Speaker Diarization Using Riemannian Manifold Clustering**

Che-Wei Huang, Bo Xiao, Panayiotis G. Georgiou, Shrikanth S. Narayanan; University of Southern California, USA

We address the problem of speaker clustering for robust unsupervised speaker diarization. We model each speaker-homogeneous segment as one single full multivariate Gaussian probability density function (pdf) and take into consideration the Riemannian property of Gaussian pdfs. By assuming that segments from different speakers lie on different (possibly intersected) sub-manifolds of the manifold of Gaussian pdfs, we formulate the original problem as a Riemannian manifold clustering problem. To apply the computationally simple Riemannian locally linear embedding (LLE) algorithm, we impose a constraint on the length of each segment so as to ensure the fitness of single-Gaussian modeling and to increase the chance that all k-nearest neighbors of a pdf are from the same sub-manifold (speaker). Experiments on the microphone-recorded conversational interviews from NIST 2010 speaker recognition evaluation set demonstrate promising results of less than 1% DER.

**Towards a Complete Binary Key System for the Speaker Diarization Task**

Héctor Delgado 1, Corinne Fredouille 2, Javier Serrano 1; 1Universidad Autónoma de Barcelona, Spain; 2LIA, France

Speaker diarization is the task of partitioning an audio stream into homogeneous segments according to speaker identity. Today state-of-the-art speaker diarization systems have achieved very competitive performance. However, any small improvement in Diarization Error Rate (DER) is usually subject to very large processing times (real time factor above one), which makes systems not suitable for some time-critical, real-life applications. Recently,
a novel fast speaker diarization technique based on speaker modeling using binary keys was presented. The proposed technique speeds up the process up to ten times faster than real-time with little increase of DER. Although the approach shows great potential, the presented results are still preliminary. The goal of this paper is to further investigate this technique, in order to move towards a complete binary key based system for the speaker diarization task. Preliminary experiments in Speech Activity Detection (SAD) based on binary keys show the feasibility of the binary key modeling approach for this task. Furthermore, the system has been tested on two different kinds of test data: meeting audio recordings and TV shows. The experiments carried out on NIST RT05 and REPERE databases show promising results and indicate that there is still room for further improvement.

An Iterative Speaker Re-Diarization Scheme for Improving Speaker-Based Entity Extraction in Multimedia Archives

Houman Ghaemmaghami, David Dean, Sridha Sridharan; Queensland University of Technology, Australia

In this paper we present a novel scheme for improving speaker diarization by making use of repeating speakers across multiple recordings within a large corpus. We call this technique speaker re-diarization and demonstrate that it is possible to reuse the initial speaker-linked diarization outputs to boost diarization accuracy within individual recordings. We first propose and evaluate two novel re-diarization techniques. We demonstrate their complementary characteristics and fuse the two techniques to successfully conduct speaker re-diarization across the SAVT-BNEWS corpus of Australian broadcast data. This corpus contains recurring speakers in various independent recordings that need to be linked across the dataset. We show that our speaker re-diarization approach can provide a relative improvement of 23% in diarization error rate (DER), over the original diarization results, as well as improve the estimated number of speakers and the cluster purity and coverage metrics.

Speaker Diarization Using Gesture and Speech

Binyam Gebrekidan Gebre 1, Peter Wittenburg 1, Sebastian Drude 1, Marjijn Huijbregts 2, Tom Heskes 2; 1MPI for Psycholinguistics, The Netherlands; 2 Radboud Universiteit Nijmegen, The Netherlands

We demonstrate how the problem of speaker diarization can be solved using both gesture and speaker parametric models. The novelty of our solution is that we approach the speaker diarization problem as a speaker recognition problem after learning speaker models from speech samples corresponding to gestures (the occurrence of gestures indicates the presence of speech and the location of gestures indicates the identity of the speaker). This new approach offers many advantages: comparable state-of-the-art performance, faster computation and more flexibility. In our implementation, parametric models are used to model speakers' voice and their gestures: more specifically, Gaussian mixture models are used to model the voice characteristics of each person and their gestures: more specifically, Gaussian mixture models are used to model speakers' vocal tract characteristics. These individual speaker characteristics are reflected in a speech signal when speakers pronounce a given phoneme. The current work hypothesizes that clusters within a phoneme spoken by multiple speakers roughly correspond to different speakers. Based on this hypothesis, a Gaussian mixture model (GMM) based phoneme background model (PBM) is estimated. The components of such a PBM are used as a set of relevance variables in information bottleneck based speaker diarization system. Experiments are done using phone transcripts obtained from ground-truth and automatic speech recognition (ASR) system to estimate the PBM. The diarization experiments done on meeting recordings from AMI and NISTRT corpora show that the proposed method achieves significant improvements over the system using a background model which ignores phoneme information.

Is Incremental Cross-Show Speaker Diarization Efficient for Processing Large Volumes of Data?

Grégor Dupuy, Sylvain Meignier, Yannick Estève; LIUM, France

Current cross-show diarization systems are mainly based on an overall clustering process which handles all the shows within a collection simultaneously. This approach has already been studied in various situations and seems to be the best way so far to achieve low error rates. However, this process has limits in realistic applicative contexts where large and dynamically increasing collections have to be processed. In this paper we investigate the use of an incremental clustering cross-show speaker diarization architecture to iteratively process new shows within an existing collection. The new shows to be inserted are processed one after another, according to the chronological order of their broadcasting dates. Experiments were conducted on the data distributed for the ETAPE and the REPERE French evaluation campaigns. It consist of 142 hours of data collected from 310 shows, from a period from Sept. 2010 to Oct. 2012.

Detecting and Labeling Speakers on Overlapping Speech Using Vector Taylor Series

Pranay Dighe, Marc Ferràs, Hervé Bourlard; Idiap Research Institute, Switzerland

Successfully modeling overlapping speech is a crucial step towards improving the performance of current speaker diarization systems. In this direction, we present ongoing work on a novel Multi-Class Vector Taylor Series (MC-VTS) approach that models overlapping speech from knowledge of the individual speaker models and the feature extraction process. We explore several variants of the MC-VTS technique that aim at modeling overlapping speech more precisely. Bootstrapping the algorithm with both oracle and diarization output segmentations, we show the potential of this approach in terms of overlapping speech detection and speaker labeling performances through a set of experiments on far-field microphone meeting data.

Phoneme Background Model for Information Bottleneck Based Speaker Diarization

Sree Harsha Yella, Petr Matlicek, Hervé Bourlard; Idiap Research Institute, Switzerland

Acoustic variability of speakers arises due to differences in their vocal tract characteristics. These individual speaker characteristics are reflected in a speech signal when speakers pronounce a given phoneme. The current work hypothesizes that clusters within a phoneme spoken by multiple speakers roughly correspond to different speakers. Based on this hypothesis, a Gaussian mixture model (GMM) based phoneme background model (PBM) is estimated. The components of such a PBM are used as a set of relevance variables in information bottleneck based speaker diarization system. Experiments are done using phone transcripts obtained from ground-truth and automatic speech recognition (ASR) system to estimate the PBM. The diarization experiments done on meeting recordings from AMI and NISTRT corpora show that the proposed method achieves significant improvements over the system using a background model which ignores phoneme information.

NOTES
Diarizing Large Corpora Using Multi-Modal Speaker Linking
Marc Ferras\textsuperscript{1}, Stefano Masneri\textsuperscript{2}, Oliver Schreer\textsuperscript{2}, Hervé Bourlard\textsuperscript{1}; \textsuperscript{1}Idiap Research Institute, Switzerland; \textsuperscript{2}Fraunhofer HHI, Germany

Speaker diarization of a collection of recordings with uniquely identified speakers is a challenging task. A system addressing such task must account for the inter-session variability present from recording to recording and it is asked to scale well to massive amounts of data. In this paper we use a two-stage approach to corpus-wide speaker diarization involving speaker diarization and speaker linking stages. The speaker linking system agglomeratively clusters speaker factor posterior distributions obtained via Joint Factor Analysis using the Ward method and the Hotteling t-square statistic as distance measure. We extend this framework to link speakers based on both speech and visual modalities to improve the robustness of the system. The system is evaluated using the data collected for the Augmented Multiparty Interaction (AMI) project, involving over one hundred meetings. We provide results in terms of within-recording and across-recording diarization error rates (DER) to support the effectiveness of multi-modal speaker linking to enable large scale speaker diarization.

Multimodal Understanding for Person Recognition in Video Broadcasts
Frederic Bechet\textsuperscript{1}, Meriem Bendris\textsuperscript{1}, Delphine Charlet\textsuperscript{1}, Géraldine Damnati\textsuperscript{2}, Benoît Favre\textsuperscript{1}, Mickael Rouvier\textsuperscript{1}, Remi Auguste\textsuperscript{3}, Benjamin Bigot\textsuperscript{4}, Richard Dufour\textsuperscript{4}, Corinne Froudouille\textsuperscript{4}, Georges Linares\textsuperscript{4}, Jean Martinet\textsuperscript{3}, Gregory Senay\textsuperscript{4}, Pierre Tirilly\textsuperscript{1}; \textsuperscript{1}LIF (UMR 7279), France; \textsuperscript{2}Orange Labs, France; \textsuperscript{3}LIFL, France; \textsuperscript{4}LIA, France

This paper describes a multi-modal person recognition system for video broadcast developed for participating in the Defi-Reper challenge. The main track of this challenge targets the identification of all persons occurring in a video either in the audio modality (speakers) or the image modality (faces). This system is developed by the PERCOL team involving 4 research labs in France and was ranked first at the 2014 Defi-Reper challenge. The main scientific issue addressed by this challenge is the combination of audio and video information extraction processes for improving the extraction performance in both modalities. In this paper, we present the strategy followed by the PERCOL team for speaker identification based on enriching the speaker diarization with features related to the "understanding" of the video scenes: text overlay transcription and analysis, automatic situation identification (TV set, report), the amount of people visible, TV set disposition and even the camera when available. Experiments on the REPERE corpus show interesting results on the speaker identification system enriched by the scene understanding features and the usefulness of the speaker to identify faces.

Notes

Oral Session 9: Robust ASR
Garnet 213-218
17:00 – 19:00, Monday 15 September 2014
Chairs: Rita Singh and Jasha Droppo

Comparing Time-Frequency Representations for Directional Derivative Features
James Gibson, Maarten Van Segbroeck, Shrikanth S. Narayanan; University of Southern California, USA

Robust Speech Recognition with Speech Enhanced Deep Neural Networks
Jun Du\textsuperscript{1}, Qing Wang\textsuperscript{1}, Tian Gao\textsuperscript{1}, Yong Xu\textsuperscript{1}, Li-Rong Dai\textsuperscript{1}, Chin-Hui Lee\textsuperscript{2}; \textsuperscript{1}USTC, China; \textsuperscript{2}Georgia Institute of Technology, USA

We propose a signal pre-processing front-end to enhance speech from the log-Mel spectrogram. By contrast, a number of techniques typically operate on the features or on the parameters of the acoustic models themselves. By contrast, a number of techniques typically operate on the features or on the parameters of the acoustic models themselves. Noise-robust automatic speech recognition (ASR) systems rely on feature and/or model compensation. Existing compensation techniques typically operate on the features or on the parameters of the acoustic models themselves. By contrast, a number of techniques typically operate on the features or on the parameters of the acoustic models themselves.
normalization techniques have been defined in the field of speaker verification that operate on the resulting log-likelihood scores. In this paper, we provide a theoretical motivation for likelihood normalization due to the so-called "hubness" phenomenon and we evaluate the benefit of several normalization techniques on ASR accuracy for the 2nd CHiME Challenge task. We show that symmetric normalization (S-norm) reduces the relative error rate by 43% alone and by 10% after feature and model compensation.

Identifying the Human-Machine Differences in Complex Binaural Scenes: What Can Be Learned from Our Auditory System

Constantin Spille, Bernd T. Meyer; Carl von Ossietzky Universität Oldenburg, Germany

Mon O-9-4

Previous comparisons of human speech recognition (HSR) and automatic speech recognition (ASR) focused on monaural signals in additive noise, and showed that HSR is far more robust against intrinsic and extrinsic sources of variation than conventional ASR. The aim of this study is to analyze the man-machine gap (and its causes) in more complex acoustic scenarios, particularly in scenes with two moving speakers, reverberation and diffuse noise. Responses of nine normal-hearing listeners are compared to errors of an ASR system that employs a binaural model for direction-of-arrival estimation and beamforming for signal enhancement. The overall man-machine gap is measured in terms for the speech recognition threshold (SRT), i.e., the signal-to-noise ratio at which a 50% recognition rate is obtained. The comparison shows that the gap amounts to 16.7 dB SRT difference which exceeds the difference of 10 dB found in monaural situations. Based on cross comparisons that use oracle knowledge (e.g., the speakers' true position), incorrect responses are attributed to localization errors (7 dB) or missing spectral information to distinguish between speakers with different gender (3 dB). The comparison hence identifies specific ASR components that can profit from learning from binaural auditory signal processing.

Robust Speech Recognition Using Long Short-Term Memory Recurrent Neural Networks for Hybrid Acoustic Modelling

Jürgen T. Geiger, Zixing Zhang, Felix Weninger, Björn Schuller, Gerhard Rigoll; Technische Universität München, Germany

Mon O-9-6

One method to achieve robust speech recognition in adverse conditions including noise and reverberation is to employ acoustic modelling techniques involving neural networks. Using long short-term memory (LSTM) recurrent neural networks proved to be efficient for this task in a setup for phoneme prediction in a multi-stream GMM-HMM framework. These networks exploit a self-learnt amount of temporal context, which makes them especially suited for a noisy speech recognition task. One shortcoming of this approach is the necessity of a GMM acoustic model in the multi-stream framework. Furthermore, potential modelling power of the network is lost when predicting phonemes, compared to the classical hybrid setup where the network predicts HMM states. In this work, we propose to use LSTM networks in a hybrid HMM setup, in order to overcome these drawbacks. Experiments are performed using the medium-vocabulary recognition track of the 2nd CHiME challenge, containing speech utterances in a reverberant and noisy environment. A comparison of different network topologies for phoneme or state prediction used either in the hybrid or double-stream setup shows that state prediction networks perform better than networks predicting phonemes, leading to state-of-the-art results for this database.

Joint Adaptation and Adaptive Training of TVWR for Robust Automatic Speech Recognition

Shilin Liu, Khe Chai Sim; National University of Singapore, Singapore

Mon O-9-6

Context-dependent Deep Neural Network has obtained consistent and significant improvements over the Gaussian Mixture Model (GMM) based systems for various speech recognition tasks. However, since DNN is discriminatively trained, it is more sensitive to label errors and is not reliable for unsupervised adaptation. Moreover, DNN parameters do not have a clear and meaningful interpretation, therefore, it has been difficult to develop effective adaptation techniques for the DNNs. Nevertheless, unadapted multi-style trained DNNs have already shown superior performance to the GMM system with joint noise/speaker adaptation and adaptive training. Recently, Temporally Varying Weight Regression (TVWR) has been successfully applied to combine DNN and GMM for robust unsupervised speaker adaptation. In this paper, joint speaker/noise adaptation and adaptive training of TVWR using DNN posteriors are investigated for robust speech recognition. Experimental results on the Aurora 4 corpus showed that after joint adaptation and adaptive training, TVWR achieved 21.3% and 11.6% relative improvements over the DNN baseline system and the best system in currently reported literatures, respectively.

Oral Session 10: Implementation of Language Model Algorithms

Period 202-203
17:00 – 19:00, Monday 15 September 2014
Chairs: Murat Saraclar and Seiichi Nakagawa

Efficient GPU-Based Training of Recurrent Neural Network Language Models Using Spliced Sentence Bunch

X. Chen, Y. Wang, X. Liu, Mark J.F. Gales, Philip C. Woodland; University of Cambridge, UK

Mon O-10-1

Recurrent neural network language models (RNNLMs) are becoming increasingly popular for a range of applications including speech recognition. However, an important issue that limits the quantity of data, and hence their possible application areas, is the computational cost in training. A standard approach to handle this problem is to use class-based outputs, allowing systems to be trained on CPUs. This paper describes an alternative approach that allows RNNLMs to be efficiently trained on GPUs. This enables larger quantities of data to be used, and networks with an unclustered, full output layer to be trained. To improve efficiency on GPUs, multiple sentences are “spliced” together for each mini-batch or “bunch” in training. On a large vocabulary conversational telephone speech recognition task, the training time was reduced by a factor of 27 over the standard CPU-based RNNLM toolkit. The use of an unclustered, full output layer also improves perplexity and recognition performance over class-based RNNLMs.

NOTES
Word Pair Approximation for More Efficient Decoding with High-Order Language Models
David Nolden, Ralf Schlüter, Hermann Ney; RWTH Aachen University, Germany
Mon-O-102

The search effort in LVCSR depends on the order of the language model (LM); search hypotheses are only recombined once the LM allows for it. In this work we show how the LM dependence can be partially eliminated by exploiting the well-known word pair approximation. We enforce preemptive unigram- or bigram-like LM recombination at word boundaries. We capture the recombination in a lattice, and later expand the lattice using LM rescoring. LM rescoring unfolds the same search space which would have been encountered without the preemptive recombination, but the overall efficiency is improved, because the amount of redundant HMM expansion in different LM contexts is reduced. Additionally, we show how to expand the recombined hypotheses on-the-fly, omitting the intermediate lattice form. Our new approach allows using the full n-gram LM for decoding, but based on a compact unigram- or bigram search space. We show that our approach works better than common lattice rescoring pipelines, where a pruned lower-order LM is used to generate lattices; such pipelines suffer from the weak lower-order LM, which guides the pruning sub-optimally. Our new decoding approach improves the runtime efficiency by up to 40% at equal precision when using a large vocabulary and high-order LM.

Comparing Approaches to Convert Recurrent Neural Networks into Backoff Language Models for Efficient Decoding
Heike Adel1, Katrin Kirchhoff2, Ngoc Thang Vu1, Dominic Telaar1, Tanja Schultz1; 1KIT, Germany; 2University of Washington, USA
Mon-O-103

In this paper, we investigate and compare three different possibilities to convert recurrent neural network language models (RNNLMs) into backoff language models (BNLM). While RNNLMs often outperform traditional n-gram approaches in the task of language modeling, their computational demands make them unsuitable for an efficient usage during decoding in an LVCSR system. It is, therefore, of interest to convert them into BNLMs in order to integrate their information into the decoding process. This paper compares three different approaches: a text based conversion, a probability based conversion and an iterative conversion. The resulting language models are evaluated in terms of perplexity and mixed error rate in the context of the Code-Switching data corpus SEAME. Although the best results are obtained by combining the results of all three approaches, the text based conversion approach alone leads to significant improvements on the SEAME corpus as well while offering the highest computational efficiency. In total, the perplexity can be reduced by 11.4% relative on the evaluation set and the mixed error rate by 3.0% relative on the same data set.

Removing Redundancy from Lattices
David Nolden1, Hagen Soltau2, Daniel Povey3, Pegah Gahremani3, Lidia Mangu2, Hermann Ney1; 1RWTH Aachen University, Germany; 2IBM T.J. Watson Research Center, USA; 3Johns Hopkins University, USA
Mon-O-104

In this work we present a novel word lattice filtering algorithm which removes redundancy from lattices. We use the filtering algorithm to analyze lattices obtained from dynamic network and transducer-based LVCSR decoders from several sites regarding size, coverage, and redundancy. We show that our filtering algorithm reduces the size of lattices by 30 to 90% without degrading the oracle word error rate.

Lattice Decoding and Rescoring with Long-Span Neural Network Language Models
Martin Sundermeyer, Zoltan Tüske, Ralf Schlüter, Hermann Ney; RWTH Aachen University, Germany
Mon-O-105

With long-span neural network language models, considerable improvements have been obtained in speech recognition. However, it is difficult to apply these models if the underlying search space is large.

In this paper, we combine previous work on lattice decoding with long short-term memory (LSTM) neural network language models. By adding refined pruning techniques, we are able to reduce the search effort by a factor of three.

Furthermore, we introduce two novel approximations for full lattice rescoring, which opens the potential of lattice-based speech recognition techniques. Compared to 1000-best lists, we find that we can increase the word error rate improvements obtained with LSTMs from 8.2% to 10.7% relative over a state-of-the-art baseline, while the resulting lattices are even considerably smaller. In addition, we investigate the use of LSTMs for Babel Assamese keyword search, obtaining significant improvements of 2.5% relative.

Word-Phrase-Entity Language Models: Getting More Mileage out of N-Grams
Michael Levit, Sarangarajan Parthasarathy, Shuangyu Chang, Andreas Stolcke, Benoît Dumoulin; Microsoft, USA
Mon-O-106

We present a modification of the traditional n-gram language modeling approach that departs from the word-level data representation and seeks to re-express the training text in terms of tokens that could be either words, common phrases or instances of one or several classes. Our iterative optimization algorithm considers alternative parses of the corpus in terms of these tokens, re-estimates token n-gram probabilities and also updates the oracle word error rate.

Oral Session 11: Speaker Recognition — Noise and Channel Robustness
Peridot 204-205
17:00 – 19:00, Monday 15 September 2014
Chairs: Niko Brummer and Jan “Honza” Černocký

A Novel Boosting Algorithm for Improved i-Vector Based Speaker Verification in Noisy Environments
Sourjya Sarkar, K. Sreenivasa Rao; IIT Kharagpur, India
Mon-O-111

This paper explores the significance of an ensemble of boosted Support Vector Machine (SVM) classifiers in the i-vector framework.

NOTES
for speaker verification (SV) in noisy environments. Prior work in this field have established the significance of supervector-based approaches and more specifically the i-vector extraction paradigm for robust SV. However, in highly degraded environments, SVMs trained using i-vectors are susceptible to misclassifications. For enhanced classification accuracy, we explore the impact of multiple SVM classifiers trained by adaptive boosting. To mitigate the effect of statistical mismatches due to difference in utterance lengths and data imbalance caused by a disproportionate ratio of target speaker and impostor utterances, we propose a novel combination scheme of the adaptive boosting algorithm with a data generation technique using partitioned utterances. All experiments are conducted on the NIST-SRE-2003 database under mismatched conditions with training utterances degraded by 4 types of additive noises (car, factory, pink and white) collected from the NOISEX-92 database, at 0 dB and 5 dB SNRs. Results indicate that the proposed method significantly outperforms the baseline i-vector SVM based SV systems across all noisy environments.

Using Deep Belief Networks for Vector-Based Speaker Recognition

W.M. Campbell; MIT Lincoln Laboratory, USA

Deep belief networks (DBNs) have become a successful approach for acoustic modeling in speech recognition. DBNs exhibit strong approximation properties, improved performance, and are parameter efficient. In this work, we propose methods for applying DBNs to speaker recognition. In contrast to prior work, our approach to DBNs for speaker recognition starts at the acoustic modeling layer. We use sparse-output DBNs trained with both unsupervised and supervised methods to generate statistics for use in standard vector-based speaker recognition methods. We show that a DBN can replace a GMM UBM in this processing. Methods, qualitative analysis, and results are given on a NIST SRE 2012 task. Overall, our results show that DBNs show competitive performance to modern approaches in an initial implementation of our framework.

A Deep Neural Network Speaker Verification System Targeting Microphone Speech

Yun Lei, Luciana Ferrer, Mitchell McLaren, Nicolas Scheffer; SRI International, USA

We recently proposed the use of deep neural networks (DNN) in place of Gaussian Mixture models (GMM) in the i-vector extraction process for speaker recognition. We have shown significant accuracy improvements on the 2012 NIST speaker recognition evaluation (SRE) telephone conditions. This paper explores how this framework can be effectively used on the microphone speech conditions of the 2012 NIST SRE. In this new framework, the verification performance greatly depends on the data used for training the DNN. We show that training the DNN using both microphone and microphone speech data can yield significant improvements. An in-depth analysis of the influence of telephone speech data on the microphone conditions is also shown for both the DNN and GMM systems. We conclude by showing that the the GMM system is always outperformed by the DNN system on the telephone-only and microphone-only conditions, and that the new DNN / i-vector framework can be successfully used providing a good match in the training data.

Application of Convolutional Neural Networks to Speaker Recognition in Noisy Conditions

Mitchell McLaren, Yun Lei, Nicolas Scheffer, Luciana Ferrer; SRI International, USA

This paper applies a convolutional neural network (CNN) trained for automatic speech recognition (ASR) to the task of speaker identification (SID). In the CNN/i-vector front end, the sufficient statistics are collected based on the outputs of the CNN as opposed to the traditional universal background model (UBM). Evaluated on heavily degraded speech data, the CNN/i-vector front end provides performance comparable to the UBM/i-vector baseline. The combination of these approaches, however, is shown to provide improvements of 26% in miss rate to considerably outperform the fusion of two different features in the traditional UBM/i-vectors approach. An analysis of the language- and channel-dependency of the CNN/i-vector approach is also provided to highlight future research directions.

SVM Based Speaker Recognition: Harnessing Trials with Multiple Enrollment Sessions

Jason Pelecanos1, Weizhong Zhu1, Sibel Yaman2; 1IBM T.J. Watson Research Center, USA; 2Apple, USA

In this paper we extend a variation of the trial-based SVM speaker verification work proposed by Cumani et al to exploit multiple enrollment sessions. Specifically, Cumani proposed the use of a 2nd order SVM kernel for the binary classification of basic trials. In this new work, trials with multiple enrollment sessions are modelled by stacking the i-vectors of the test and enrollment sessions. We further exploit the fact that the score should be independent of the enrollment recording order and present a simplified 2nd order polynomial kernel scoring function accordingly.

In the second part of this work we examine the utility of enrollment pruning for multi-session enrollments. Past work demonstrates that pruning can be beneficial for PLDA based systems. We examine the effects of enrollment pruning in the context of the proposed SVM model.

The results demonstrate that the multi-session enrollment SVM kernel is generally better than the model trained using single sessions. The model is also comparable in performance to the PLDA based approach. Further gains are observed through combination of the PLDA and SVM scores.

I-Vector Speaker Verification Based on Phonetic Information Under Transmission Channel Effects

Laura Fernández Gallardo1, Michael Wagner1, Sebastian Möller2; 1University of Canberra, Australia; 2T-Labs, Germany

Past studies have shown evidence of important speaker-specific content in the higher frequencies of the speech, which are filtered out by narrowband channels. Besides, wideband transmissions, which are gaining ground over narrowband communications, offer an extended range of frequencies which account not only for better speech quality and intelligibility, but also for an improved speaker recognition performance. In this work, different phoneme classes (fricatives, nasals, and vowels) were removed from speech of different bandwidths, and a series of i-vector based speaker verification experiments were conducted. Our results show that the performance enhancement with clean wideband speech with respect to clean narrowband speech is principally due to the
A Real-Time MRI Study of Articulatory Setting in Second Language Speech

Andrés Benitez, Vikram Ramanarayanan, Louis Goldstein, Shrikanth S. Narayanan; University of Southern California, USA

Previous work has shown that languages differ in their articulatory setting, the postural configuration that the vocal tract articulators tend to adopt when they are not engaged in any active speech gesture, and that this posture might be specified as part of the phonological knowledge speakers have of the language. This study tests whether the articulatory setting of a language can be acquired by non-native speakers. Three native speakers of German who had learned English as a second language were imaged using real-time MRI of the vocal tract while reading passages in German and English, and features that capture vocal tract posture were extracted from the inter-speech pauses in their native and non-native languages. Results show that the speakers exhibit distinct inter-speech postures in each language, with a lower and more retracted tongue in English, consistent with classic descriptions of the differences between the German and the English articulatory settings. This supports the view that non-native speakers may acquire relevant features of the articulatory setting of a second language, and also lends further support to the idea that articulatory setting is part of a speaker’s phonological competence in a language.

Retroflex and Bunched English /r/ with Physical Models of the Human Vocal Tract

Takayuki Araii; Sophia University, Japan

It is known that American English /r/ can be produced as a retroflex or bunched /r/, but it can be challenging to teach students how to articulate both. We already developed a physical model for retroflex /r/ and demonstrated that the model produces the /r/ sound. However, almost no studies have reported a physical model for bunched /r/. We developed a new physical model using sliding blocks for the lips and tongue to help teach students how to produce bunched /r/. We recorded several sets of sounds produced by the models, analyzed the output signals, and used them for perceptual experiments. Acoustic analysis and perceptual experiments confirmed that the retroflex and bunched /r/ models produced clear American /r/ sounds, and that the narrow constriction placed between 5-7 cm from the lips seems to be the key in producing these sounds. Furthermore, bunched /r/ with lip rounding produced the most clear /r/ sound. Both models are helpful for practicing pronunciation because learners can readily see there are two ways to produce /r/, they can see and alter the tongue position manually, and they can hear the output sounds.

Parameterization of Articulatory Pattern in Speakers with ALS

Panying Rong¹, Yana Yunusova², James D. Berry³, Lorne Zinman², Jordan R. Green¹; ¹MGH Institute of Health Professions, USA; ²University of Toronto, Canada; ³Massachusetts General Hospital, USA

A combination of parallel factor analysis (PARAFAC) and principal component analysis (PCA) was used to parameterize the articulatory pattern of tongue, jaw and lip movements in 8 English vowels produced by 7 subjects with amyotrophic lateral sclerosis (ALS). A two-factor PARAFAC model derived an overall articulatory pattern represented by two basic modes dominated by tongue raising and advancement, respectively. The relation between the two articulatory modes and the acoustic formants (F1, F2) followed a simple one-to-one linear mapping. The PCA on the residuals of the PARAFAC model showed various individualized articulatory features superimposed on the overall pattern. These articulatory features contributed in a systematic way to the acoustic deviation across different subjects. The parameterization approach (1) provided a simple and generalizable way to explore the underlying articulatory mechanism of speech decline in ALS and (2) accounted for the articulatory features across affected individuals. With further development of the approach and a comparison with the articulatory pattern for healthy subjects, it is possible to derive a set of quantitative articulatory indicators of speech impairment in ALS.

Missing Samples Estimation in Electromagnetic Articulography Data Using Equality Constrained Kalman Smoother

Sujith P., Prasanta Kumar Ghosh; Indian Institute of Science, India

Electromagnetic articulography (EMA) data provides the movement of sensors attached to different articulators of a subject when the subject is speaking. EMA data often contains missing segments due to sensor failure. In this work, we propose an equality constrained Kalman smoother to estimate the missing samples in the EMA data. We incorporate the dynamics of the articulatory movement for missing samples estimation by considering the EMA data vector as the observations from a linear dynamical system. The proposed approach gives 41% reduction on the root mean square error of the estimates compared to the minimum mean square error estimator which does not utilize the dynamics of the articulatory movement. When compared to the maximum a-posteriori estimation with continuity constraints (MAPC) which incorporates smoothness of the articulatory trajectory during estimation, the proposed approach gives an average performance improvement of 4.6%.

Palate-Referenced Articulatory Features for Acoustic-to-Articulator Inversion

An Ji, Michael T. Johnson, Jeff Berry; Marquette University, USA

The selection of effective articulatory features is an important component of tasks such as acoustic-to-articulator inversion and articulatory synthesis. Although it is common to use direct articulatory sensor measurements as feature variables, this approach fails to incorporate important physiological information such as palate height and shape and thus is not as representative of vocal tract cross section as desired. We introduce a set of articulator feature variables that are palate referenced and normalized with respect

NOTES
to the articulatory working space in order to improve the quality of
the vocal tract representation. These features include normalized
horizontal positions plus the normalized palatal height of two
midpalatal and one lateral tongue sensor, as well as normalized
lip separation and lip protrusion. The quality of the feature repre-
sentation is evaluated subjectively by comparing the variances and
vowel separation in the working space and quantitatively through
measurement of acoustic-to-articulator inversion error. Results
indicate that the palate-referenced features have reduced variance
and increased separation between vowels spaces and substantially
lower inversion error than direct sensor measures.

A Study on the Improvement of Measurement
Accuracy of the Three-Dimensional Electromagnetic
Articulography
Hidetsugu Uchida, Kohei Wakamiya, Tokihiko
Kaburagi; Kyushu University, Japan
Mon SP 2b-2
The alignment of the transmitter coils for the three-dimensional
electromagnetic articulography (3D-EMA), an instrument used to
measure articulatory movements, was studied. The receiver coils
of the 3D-ENA are used as position markers and are placed in
an alternating magnetic field produced by multiple transmitter
coils. The estimation of state (the position and orientation) of
each receiver coil is based on the minimization of signal error
between the measured and predicted receiver signals using a
model of the magnetic field. Previous studies report a noticeable
increase in the position estimation error at a specific portion of the
measurement region irrespective of small signal error values. The
existence of non-uniqueness in the position estimation problem
is hypothesized to be the cause of this problem. To resolve the
problem, we optimized the alignment of the transmitter coils by
maximizing the difference between the receiver signals at any two
states in the measurement region and evaluated the alignment
using a computer simulation and an experiment. As a result, a
measurement accuracy of approximately 0.4 mm was obtained.

Special Session 2(b): INTERSPEECH 2014
Computational Paralinguistics Challenge (ComParE) II
Peridot 206
17:00 – 19:00, Monday 15 September 2014
Chairs: Anton Batliner and Stefan Steidl

High-Level Speech Event Analysis for Cognitive
Load Classification
Claude Montacié, Marie-José Caraty; STIH (EA 4509),
France
Mon SP 2b-1
The Cognitive Load (CL) refers to the load imposed on an individ-
ual’s cognitive system when performing a given task, and is usually
associated with the limitations of the human working memory.
Stress, fatigue, lower ability to make decisions and perceptual
narrowing are induced by cognitive overload which occurs when
too much information has to be processed. As many physiological
measures and for a noninvasive measurement, speech features
have been investigated in order to find reliable indicators of CL
evels. In this paper, we have investigated high-level speech events
automatically detected using the CMU-Sphinx toolkit for speech
recognition. Temporal events (speech onset latency, event starting
time-codes, pause and phone segments) were extracted from the
speech transcriptions (phoneme, word, silent pause, filled pause,
breathing). Seven audio feature sets related to the speech events
were designed and assessed. Three-class SVM classifiers (Low,
Medium and High level) were developed and assessed on the
CSLE (Cognitive-Load with Speech and EGG) databases provided
for the Interspeech 2014 Computational Paralinguistics Challenge.
These experiments have shown an improvement of 1.5% on the Test set
compared to the official baseline Unweighted Average Recall (UAR).

On the Use of Bhattacharyya Based GMM Distance
and Neural Net Features for Identification of
Cognitive Load Levels
Tin Lay Nwe, Trung Hieu Nguyen, Bin Ma; A*STAR,
Singapore
Mon SP 2b-2
This paper presents a method for detecting cognitive load levels
from speech. When speech is modulated by different levels of
cognitive load, acoustic characteristics of speech change. In this
paper, we measure acoustic distance of a stressed utterance from the
baseline stress free speech using GMM-SVM kernel with
Bhattacharyya based GMM distance. In addition, it is believed that
airflow structure of speech production is nonlinear. This motivates
us to investigate better techniques to capture nonlinear charac-
teristic of stress information in acoustic features. Inspired by the
recent success of neural networks for representation learning, we
employ a single hidden layer feed forward network with non-linear
activation to extract the feature vectors. Furthermore, people have
different reactions to a particular task load. This inter-speaker
difference in stress responses presents a major challenge for stress
level detection. We use a bootstrapped training process to learn the
stress response of a particular speaker. We perform experiments
using data sets from Cognitive Load with Speech and EGG (CLSE)
provided for the Cognitive Load Sub-Challenge of the INTERSPEECH
2014 Computational Paralinguistics Challenge. The results show
that the system with our proposed strategies performs well on
validation and test sets.

Prediction of Cognitive Load from Speech with the
VOQAL Voice Quality Toolbox for the InterSpeech
2014 Computational Paralinguistics Challenge
Mark Huckvale; University College London, UK
Mon SP 2b-3
This paper describes the UCL system for the cognitive load task of
the Interspeech 2014 Computational Paralinguistics Challenge. The
UCL system evaluates whether additional voice features computed
by the VOQAL voice analysis toolbox improves performance over
the baseline feature set. 144 different system configurations are
evaluated on the development test set, with some systems achiev-
ing 100% classification accuracy of cognitive load in the two Stroop
subtasks. The difficulty of the reading span sub-task is shown to be
caused in part by the duration of the audio material. Performance
of the best systems on the test set confirm the importance of
building speaker dependent systems. While the VOQAL augmented
features gave the best performance on the development test set,
no benefit was found for the test set.

The UNSW Submission to INTERSPEECH 2014
ComParE Cognitive Load Challenge
Jia Min Karen Kua, Vidhyasaharan Sethu, Phu Le,
Elilathamby Ambikairajah; University of New South
Wales, Australia
Mon SP 2b-4
Speech based cognitive load estimation is a new field of research.
Due to this relative ‘lack of maturity’, a single best approach to

NOTES
building cognitive load estimation systems has not been established yet. The primary aim of this submission is to report the performance of various basic utterance level classification frameworks developed using important elements of state-of-the-art speaker recognition systems. This may lead to a suitable basis for future cognitive load estimation systems. As a consequence of being a part of a challenge, it is expected that these frameworks will be compared to a much larger number of alternative approaches than what would otherwise be possible. In keeping with this focused aim, the GMM supervector approaches along with some variants are utilised. The systems outlined in this paper include a frame-level MFCC-GMM system along with utterance level GMM-supervisor-SVM, GMM-i-vector-SVM and GMM-JFA-SVM systems. The best combined system has an accuracy (UAR) of 66.6% as evaluated on the challenge development set and 63.7% as evaluated on the test set.

Classification of Cognitive Load from Speech Using an i-Vector Framework

Maarten Van Segbroeck, Ruchir Travadi, Colin Vaz, Jangwon Kim, Matthew P. Black, Alexandros Potamianos, Shrikanth S. Narayanan; University of Southern California, USA
Mon-P-7-1, Poster

The goal in this work is to automatically classify speakers’ level of cognitive load (low, medium, high) from a standard battery of reading tasks requiring varying levels of working memory. This is a challenging machine learning problem because of the inherent difficulty in defining/measuring cognitive load and due to intra-/inter-speaker differences in how their effects are manifested in behavioral cues. We experimented with a number of static and dynamic features extracted directly from the audio signal (prosodic, spectral, voice quality) and from automatic speech recognition hypotheses (lexical information, speaking rate). Our approach to classification addressed the wide variability and heterogeneity through speaker normalization and by adopting an i-vector framework that affords a systematic way to factorize the multiple sources of variability.

Poster Session 7: Speech Synthesis I
Max Atria Gallery
17:00 – 19:00, Monday 15 September 2014
Chair: Tomoki Toda

Using Conditional Random Fields to Predict Focus Word Pair in Spontaneous Spoken English
Xiao Zang, Zhiyong Wu, Helen Meng, Jia Jia, Lianhong Cai; Tsinghua University, China
Mon-P-7-1, Poster

This paper addresses the problem of automatically labeling focus word pairs in spontaneous spoken English, where a focus word pair refers to salient part of text or speech and the word motivating it. The prediction of focus word pairs is important for speech applications such as expressive text-to-speech (TTS) synthesis and speech recognition. It can also help in better textual and intention understanding for spoken dialog systems. Traditional approaches such as support vector machines (SVMs) prediction neglect the dependency between words and meet the obstacle of the imbalanced distribution of positive and negative samples of dataset. This paper introduces conditional random fields (CRFs) to the task of automatically predicting focus word pair from lexical, syntactic and semantic features. Furthermore, several new features related to syntactic and semantic information are proposed to achieve better performance. Experiments on the publicly available Switchboard corpus demonstrate that CRF model outperforms the baseline and SVM model for focus word pair prediction, and newly proposed features can further improve performance for CRF based predictor. Specifically, compared to the low recall rate of 11.31% achieved by the SVM model, the proposed CRF based predictor can yield a high recall rate of 70.88% with little impact on precision.

Applications of Maximum Entropy Rankers to Problems in Spoken Language Processing
Richard Sproat, Keith Hall; Google, USA
Mon-P-7-2, Poster

We report on two applications of Maximum Entropy-based ranking models to problems of relevance to automatic speech recognition and text-to-speech synthesis. The first is stress prediction in Russian, a language with notoriously complex morphology and stress rules. The second is the classification of alphabetic non-standard words, which may be read as words (NATO), as letter sequences USA, or as a mixed (mymsm). For this second task we report results on English, and five other European languages.

Text-To-Speech with Cross-Lingual Neural Network-Based Grapheme-To-Phoneme Models
Xavi Gonzalvo, Monika Podsiadlo; Google, UK
Mon-P-7-3, Poster

Modern Text-To-Speech (TTS) systems need to increasingly deal with multilingual input. Navigation, social and news are all domains with a large proportion of foreign words. However, when typical monolingual TTS voices are used, the synthesis quality on such input is markedly lower. This is because traditional TTS derives pronunciations from a lexicon or a Grapheme-To-Phoneme (G2P) model which was built using a pre-defined sound inventory and a phonotactic grammar for one language only. G2P models perform poorly on foreign words, while manual lexicon development is labour-intensive, expensive and requires extra storage. Furthermore, large phoneme inventories and phonotactic grammars contribute to data sparsity in unit selection systems. We present an automatic system for deriving pronunciations for foreign words that utilises the monolingual voice design and can rapidly scale to many languages. The proposed system, based on a neural network cross-lingual G2P model, does not increase the size of the voice database, doesn’t require large data annotation efforts, is designed not to increase data sparsity in the voice, and can be sized to suit embedded applications.

Transform Mapping Using Shared Decision Tree Context Clustering for HMM-Based Cross-Lingual Speech Synthesis
Daiki Nagahama1, Takashi Nose2, Tomoki Kory Yamashita1, Takao Kobayashi1; 1Tokyo Institute of Technology, Japan; 2Tohoku University, Japan
Mon-P-7-4, Poster

This paper proposes a novel transform mapping technique based on shared decision tree context clustering (STC) for HMM-based cross-lingual speech synthesis. In the conventional cross-lingual speaker adaptation based on state mapping, the adaptation performance is not always satisfactory when there are mismatches of languages and speakers between the average voice model of input and output languages. In the proposed technique, we alleviate the effect of the mismatches on the transform mapping by introducing a language-independent decision tree constructed by STC, and

Notes
represent the average voice models using language-independent parameters are derived by LP analysis. Synthesis is carried by and dependent tree structures. We also use a bilingual speech using mixed excitation source consisting of a sequence of impulses corpus for keeping speaker characteristics between the average for voiced regions and white-noise source for unvoiced regions. voice models of different languages. The experimental results show that the proposed technique decreases both spectral and that proposed approach is may be due to the novel chaotic mixed distortions between original and generated parameter excitation source of excitation. trajectories and significantly improves the naturalness of synthetic speech while keeping the speaker similarity compared to the state mapping.

Cross-Lingual Voice Conversion-Based Polyglot Speech Synthesizer for Indian Languages Ramani B., Actlin Jeeva M.P., Vijayalakshmi P., Nagarajan T.; SSN, India

A polyglot speech synthesizer, synthesizes speech for any given monolingual or multilingual text, in a single speaker's voice. In this regard, a polyglot speech corpus is required. It is difficult to find a speaker proficient in multiple languages. Therefore, in the current work, by exploiting the acoustic similarity of phonemes across Indian languages, a polyglot speech corpus is obtained for four Indian languages and Indian English, using GMM-based cross-lingual voice conversion. The optimum target speaker and GMM topology is chosen based on the performance of a speaker identification system. It is observed that, the language that shares the most number of phonemes with the other languages, serves as the best target. A polyglot speech corpus derived in this target speaker's voice, is further used to develop an HMM-based polyglot speech synthesizer. The performance of this synthesizer is evaluated in terms of speaker identity using ABX listening test, quality using mean opinion score (MOS) and speaker switching using subjective listening test.

An Investigation of the Application of Dynamic Sinusoidal Models to Statistical Parametric Speech Synthesis Qiong Hu 1, Yannis Stylianou 2, Ranniery Maia 2, Korin Richmond 1, Junichi Yamagishi 1, Javier Latorre 2; 1University of Edinburgh, UK; 2Toshiba Research Europe, UK

This paper applies a dynamic sinusoidal synthesis model to statistical parametric speech synthesis (HTS). For this, we utilise regularised cepstral coefficients to represent both the static amplitude and dynamic slope of selected sinusoids for statistical modelling. During synthesis, a dynamic sinusoidal model is used to reconstruct speech. A preference test is conducted to compare the selection of different sinusoids for cepstral representation. Our results show that when integrated with HTS, a relatively small number of sinusoids selected according to a perceptual criterion can produce quality comparable to using all harmonics. A Mean Opinion Score (MOS) test shows that our proposed statistical system is preferred to one using mel-cepstra from pitch synchronous spectral analysis.

Chaotic Mixed Excitation Source for Speech Synthesis Hemant A. Patil, Tanvina B. Patel; DA-IICT, India

Linear Prediction (LP) analysis has proven to be very powerful and widely used method in speech analysis and synthesis. Synthesis by LP-based approach is carried by exciting an all-pole model (whose
Posters & Panels

Mon-P-8-1, Poster

Improving Language-Universal Feature Extraction with Deep Maxout and Convolutional Neural Networks

Yajie Miao, Florian Metze; Carnegie Mellon University, USA

When deployed in automated speech recognition (ASR), deep neural networks (DNNs) can be treated as a complex feature extractor plus a simple linear classifier. Previous work has investigated the utility of multilingual DNNs acting as language-universal feature extractors (LUFEs). In this paper, we explore different strategies to further improve LUFEs. First, we replace the standard sigmoid non-linearity with the recently proposed maxout units. The resulting maxout LUFEs have the nice property of generating sparse feature representations. Second, the convolutional neural network (CNN) architecture is applied to obtain more invariant feature space. We evaluate the performance of LUFEs on a cross-language ASR task. Each of the proposed techniques results in word error rate reduction compared with the existing DNN-based LUFEs. Combining the two methods together brings additional improvement on the target language.

Mon-P-8-2, Poster

Exploiting Vocal-Source Features to Improve ASR Accuracy for Low-Resource Languages

Raul Fernandez¹, Jia Cui¹, Andrew Rosenberg², Bhuvana Ramabhadran¹, Xiaodong Cui¹; ¹IBM T.J. Watson Research Center, USA; ²CUNY Queens College, USA

A traditional framework in speech production describes the output speech as an interaction between a source excitation and a vocal-tract configured by the speaker to impart segmental characteristics. In general, this simplification has led to approaches where systems that focus on phonetic segment tasks (e.g. speech recognition) make use of a front-end that extracts features that aim to distinguish between different vocal-tract configurations. The excitation signal, on the other hand, has received more attention for speaker-characterization tasks. In this work we look at augmenting the front-end in a recognition system with vocal-source features, motivated by our work with languages that are low in resources and whose phonology and phonetics suggest the need for complementary approaches to classical ASR features. We demonstrate that the additional use of such features provides improvements over a state-of-the-art system for low-resource languages from the babel program.

Mon-P-8-3, Poster

Data Augmentation for Low Resource Languages

Anton Ragni, Kate M. Knill, Shakti P. Rath, Mark J.F. Gales; University of Cambridge, UK

Recently there has been interest in the approaches for training speech recognition systems for languages with limited resources. Under the IARPA Babel program such resources have been provided for a range of languages to support this research area. This paper examines a particular form of approach, data augmentation, that can be applied to these situations. Data augmentation schemes aim to increase the quantity of data available to train the system, for example semi-supervised training, multilingual processing, acoustic data perturbation and speech synthesis. To date the majority of work has considered individual data augmentation schemes, with few consistent performance contrasts or examination of whether the schemes are complementary. In this work two data augmentation schemes, semi-supervised training and vocal tract length perturbation, are examined and combined on the Babel limited language pack configuration. Here only about 10 hours of transcribed acoustic data are available. Two languages are examined, Assamese and Zulu, which were found to be the most challenging of the Babel languages released for the 2014 Evaluation. For both languages consistent speech recognition performance gains can be obtained using these augmentation schemes. Furthermore the impact of these performance gains on a down-stream keyword spotting task are also described.

Monday, 15 September 2014, 17:00 – 19:00
Max Atria Gallery
Chair: Penny Karanasou

17:00 – 17:30
Max Atria Gallery
Poster Session 8: Multi-Lingual Cross-Lingual and Low-Resource ASR
Max Atria Gallery
17:00 – 19:00, Monday 15 September 2014
Chair: Penny Karanasou

Mon-P-8-4, Poster

About Combining Forward and Backward-Based Decoders for Selecting Data for Unsupervised Training of Acoustic Models

Denis Jouvet, Dominique Fohr; LORIA, France

This paper introduces the combination of speech decoders for selecting automatically transcribed speech data for unsupervised training or adaptation of acoustic models. Here, the combination relies on the use of a forward-based and a backward-based decoder. Best performance is achieved when selecting automatically transcribed data (speech segments) that have the same word hypotheses when processed by the Sphinx forward-based and the Julius backward-based transcription systems, and this selection process outperforms confidence measure based selection. Results are reported and discussed for adaptation and for full training from scratch, using data resulting from various selection processes, whether alone or in addition to the baseline manually transcribed data. Overall, selecting automatically transcribed speech segments that have the same word hypotheses when processed by the Sphinx forward-based and Julius backward-based recognizers, and adding this automatically transcribed and selected data to the manually transcribed data leads to significant word error rate reductions on the ESTER2 data when compared to the baseline system trained only on manually transcribed speech data.

Mon-P-8-5, Poster

Combination of Multilingual and Semi-Supervised Training for Under-Resourced Languages

František Grézl, Martin Karafiát; Brno University of Technology, Czech Republic

Multilingual training of neural networks for ASR is widely studied these days. It has been shown that languages with little training data can benefit largely from the multilingual resources for training. The use of unlabeled data for the neural network training in semi-supervised manner has also improved the ASR system performance. Here, the combination of both methods is presented. First, multilingual training is performed to obtain an ASR system to automatically transcribe the unlabeled data. Then, the automatically transcribed data are added. Two neural networks are trained — one from random initialization and one adapted from multilingual network — to evaluate the effect of multilingual training under presence of larger amount of training
data. Further, the CMLLR transform is applied in the middle of the stacked Bottle-Neck neural network structure. As the CMLLR rotates the features to better fit given model, we evaluated whether it is better to adapt the existing NN on the CMLLR features or if it is better to train it from random initialization. The last step in our training procedure is the fine-tuning on the original data.

**Investigating the Learning Effect of Multilingual Bottle-Neck Features for ASR**

Ngoc Thang Vu, Jochen Weiner, Tanja Schultz; KIT, Germany

Mon P-8-6, Poster

Deep neural networks (DNNs) have become state-of-the-art techniques of automatic speech recognition in the last few years. They can be used at the pre-processing level (Tandem or Bottle-Neck features) or at the acoustic model level (hybrid Hidden Markov Model/DNN). Moreover, they allow exploiting multilingual data to improve monolingual systems. This paper presents our investigation of the learning effect of neural networks in the context of multilingual Bottle-Neck features. For this, we perform a visual analysis of the output of the Bottle-Neck layer of a neural network using t-Distributed Stochastic Neighbor Embedding. Our results show that multilingual Bottle-Neck features seem to learn phoneme characteristics, such as the F1 and F2 formants which characterize different vowels, and other articulatory features, such as fricatives and nasals which characterize consonants. Furthermore, they seem to normalize language dependent variations and transfer the learned representation to unseen languages.

**Distributed Learning of Multilingual DNN Feature Extractors Using GPUs**

Yajie Miao, Hao Zhang, Florian Metze; Carnegie Mellon University, USA

Mon P-8-7, Poster

Multilingual deep neural networks (DNNs) can act as deep feature extractors and have been applied successfully to cross-language acoustic modeling. Learning these feature extractors becomes an expensive task, because of the enlarged multilingual training data and the sequential nature of stochastic gradient descent (SGD). This paper investigates strategies to accelerate the learning process over multiple GPU cards. We propose the DistModel and DistLang frameworks which distribute feature extractor learning by models and languages respectively. The time-synchronous DistModel has the nice property of tolerating infrequent model averaging. With 3 GPUs, DistModel achieves 2.6× speed-up and causes no loss on word error rates. When using DistLang, we observe better acceleration but worse recognition performance. Further evaluations are conducted to scale DistModel to more languages and GPU cards.

**Towards Better Performance with Heterogeneous Training Data in Acoustic Modeling Using Deep Neural Networks**

Yan Huang, Malcolm Slaney, Michael L. Seltzer, Yifan Gong; Microsoft, USA

Mon P-8-9, Poster

Modeling heterogeneous data sources remains a fundamental challenge of acoustic modeling in speech recognition. We call this the multi-condition problem because the speech data come from many different conditions. In this paper, we introduce the fundamental confusability problem in multi-condition learning, then discuss the problem formalization, the taxonomy, and the architectures for multi-condition learning. While the ideas presented are applicable to all classifiers, we focus our attention in this work on acoustic models based on deep neural networks (DNN). We propose four different strategies for multi-condition learning of a DNN that we refer to as a mixed-condition model, a condition-dependent model, a condition-normalizing model, and a condition-avoid model. Based on the experimental results on the voice search and short message dictation task and the Aurora 4 task, we show that the confusability introduced when modeling heterogeneous data depends on the source of acoustic distortion itself, the front-end feature extractor, and the classifier. We also demonstrate the best approach for dealing with heterogeneous data may not be to let the model sort it out blindly, even with a classifier as sophisticated as a DNN.
A Unified Approach for Underdetermined Blind Signal Separation and Source Activity Detection by Multichannel Factorial Hidden Markov Models

Takuya Higuchi, Hirofumi Takeda, Tomohiko Nakamura, Hirokazu Kameoka; University of Tokyo, Japan

This paper proposes to introduce a new model called “the multi-channel factorial hidden Markov Model (MFHMM)” for under-determined blind signal separation (BSS). For monaural source separation, one successful approach involves applying non-negative matrix factorization (NMF) to the magnitude spectrogram of a mixture signal, interpreted as a non-negative matrix. Up to now, multichannel extensions of NMF, which allow for the use of spatial information as an additional clue for source separation, have been proposed by several authors and proven to be an effective approach for underdetermined BSS. This approach is based on the assumption that an observed signal is a mixture of a limited number of source signals each of which has a static power spectral density scaled by a time-varying amplitude. However, many source signals in real world are non-stationary in nature and the variations of the spectral densities are much richer in time. Moreover, many sources including speech tend to stay inactive for some while until they switch to an active mode, implying that the total power of a source may depend on its underlying state. To reasonably characterize such a non-stationary nature of source signals, this paper proposes to extend the multichannel NMF model by modeling the transition of the set consisting of the spectral densities and the total power of each source using a hidden Markov model (HMM). By letting each HMM contain states corresponding to active and inactive modes, we will show that voice activity detection and source separation can be solved simultaneously through parameter inference of the present model. The experiment showed that the proposed algorithm provided a 7.65 dB improvement compared with the conventional multichannel NMF in terms of the signal-to-distortion ratio.

Enhancing Audio Source Separability Using Spectro-Temporal Regularization with NMF

Colin Vaz 1, Dimitrios Dimitriadis 2, Shrikanth S. Narayanan 1, 1University of Southern California, USA; 2AT&T Labs Research, USA

We propose a spectro-temporal regularization approach for NMF that accounts for a source’s spectral variability over time. The regularization terms allow NMF to adapt the spectral basis matrices optimally to reduce mismatch between the spectral characteristics of sources observed during training and encountered during separation. We first tested our algorithm on a simulated source separation task. Preliminary results show significant improvement of SAR, SDR, and SIR values over some current NMF methods. We also tested our algorithm on a speech enhancement task and were able to show a modest improvement of the PESQ scores of the recovered speech.

Discriminative NMF and its Application to Single-Channel Source Separation

Felix Weninger, Jonathan Le Roux, John R. Hershey, Shinji Watanabe; MERL, USA

The objective of single-channel source separation is to accurately recover source signals from mixtures. Non-negative matrix factorization (NMF) is a popular approach for this task, yet previous NMF approaches have not optimized directly this objective, despite some efforts in this direction. Our paper introduces discriminative training of the NMF basis functions such that, given the coefficients obtained on a mixture, a desired source is optimally recovered. We approach this optimization by generalizing the model to have separate analysis and reconstruction basis functions. This generalization frees us to optimize reconstruction objectives that incorporate the filtering step and SNR performance criteria. A novel multiplicative update algorithm is presented for the optimization of the reconstruction basis functions according to the proposed discriminative objective functions. Results on the 2nd CHiME Speech Separation and Recognition Challenge task indicate significant gains in source-to-distortion ratio with respect to sparse NMF, exemplar-based NMF, as well as a previously proposed discriminative NMF criterion.

Vocal Tract Length Estimation Based on Vowels Using a Database Consisting of 385 Speakers and a Database with MRI-Based Vocal Tract Shape Information

Hideki Kawahara 1, Tatsuya Kitamura 2, Hironori Takekoto 3, Ryuichi Nisimura 1, Tosho Irino 1;
1Wakayama University, Japan; 2Konan University, Japan; 3NICT, Japan

A highly-reproducible estimation method of vocal tract length (VTL) and text independent VTL estimation method are proposed based on a Japanese vowel database spoken by 385 male and female speakers ranging from age 6 to 56 and other vowel database with MRI-based vocal tract shape information. Proposed methods.
are based on interference-free power spectral representation and systematic suppression of biasing factors. MRI data is used to calibrate VTL estimation result to be represented in terms of physically meaningful unit. These databases are normalized based on the estimated VTL information to provide a reference template, which is used to implement a text independent VTL estimation method. A prototype system for text independent estimation of VTL is implemented using Matlab and runs faster than realtime on a PC.

A Graph-Based Gaussian Component Clustering Approach to Unsupervised Acoustic Modeling

Haipeng Wang ¹, Tan Lee ¹, Cheung-Chi Leung ², Bin Ma ², Haizhou Li ²; ¹Chinese University of Hong Kong, China; ²A*STAR, Singapore

This paper describes a new approach to unsupervised acoustic modeling, that is to build acoustic models for phoneme-like sub-word units from untranscribed speech data. The proposed approach is based on Gaussian component clustering. Initially a large set of Gaussian components are estimated from the untranscribed data. Then clustering is performed to group these Gaussian components into different clusters. Each cluster of Gaussian components forms an acoustic model for an induced sub-word unit. We have defined several similarity measures among the Gaussian components, and investigated several different graph-based clustering algorithms. Experiments on the TIMIT corpus demonstrate the effectiveness of our approach.

A Speech System for Estimating DailyWord Counts

Ali Ziaei, Abhijeet Sangwan, John H.L. Hansen; University of Texas at Dallas, USA

The ability to count the number of words spoken by an individual over long durations is important to researchers investigating language development, healthcare, education, etc. In this study, we attempt to build a speech system that can compute daily word counts using data from the Prof-Life-Log corpus. The task is challenging as typical audio files from Prof-Life-Log tend to be 8-to-16 hours long, where audio is collected continuously using the LENA device. This device is worn by the primary speaker and all his daily interactions are collected in fine detail. The recordings contain a wide variety of noise types with varying SNR (signal-to-noise ratio) including large crowd, babble, and competing secondary speakers. In this study, we develop a word-count estimation (WCE) system based on syllable detection and we use the method proposed by Wang and Narayanan as the baseline system [1]. We propose many modifications to the original algorithm to improve its effectiveness in noise. Particularly, we incorporate speech activity detection and enhancement techniques to remove non-speech from analysis and improve signal quality for superior syllable detection, respectively. We also investigate features derived from syllable detection for better word count estimation. The proposed method show significant improvement over the baseline.

Ensemble Modeling of Denoising Autoencoder for Speech Spectrum Restoration

Xugang Lu ¹, Yu Tsao ², Shigeki Matsuda ¹, Chiori Hori ¹; ¹NICT, Japan; ²Academia Sinica, Taiwan

Denoising autoencoder (DAE) is effective in restoring clean speech from noisy observations. In addition, it is easy to be stacked to a deep denoising autoencoder (DDAE) architecture to further improve the performance. In most studies, it is supposed that the DAE or DDAE can learn any complex transform functions to approximate the transform relation between noisy and clean speech. However, for large variations of speech patterns and noisy environments, the learned model is lack of focus on local transformations. In this study, we propose an ensemble modeling of DAE to learn both the global and local transform functions. In the ensemble modeling, local transform functions are learned by several DAEs using data sets obtained from unsupervised data clustering and partition. The final transform function used for speech restoration is a combination of all the learned local transform functions. Speech denoising experiments were carried out to examine the performance of the proposed method. Experimental results showed that the proposed ensemble DAE model provided superior restoration accuracy than traditional DAE models.
Decision Learning in Data Science: Where John Nash Meets Social Media

K.J. Ray Liu; University of Maryland, USA

With the increasing ubiquity and power of mobile devices, as well as the prevalence of social media, more and more activities in our daily life are being recorded, tracked, and shared, creating the notion of “social media”. Such abundant and still growing real-life data, known as “big data”, provide a tremendous research opportunity in many fields. To analyze, learn and understand such user-generated big data, machine learning has been an important tool and various machine learning algorithms have been developed. However, since the user-generated big data is the outcome of users’ decisions, actions, and their socio-economic interactions, which are highly dynamic, without considering users’ local behaviors and interests, existing learning approaches tend to focus on optimizing a global objective function at the macroeconomic level, while totally ignoring users’ local decisions at the microeconomic level. As such, there is a growing need in bridging machine/social learning with strategic decision making, which are two traditionally distinct research disciplines, to be able to jointly consider both global phenomena and local effects to understand/model/analyze better the newly arising issues in the emerging social media. In this talk, we present the notion of “decision learning” that can involve users’ behaviors and interactions by combining learning with strategic decision making. We will discuss some examples from social media with real data to show how decision learning can be used to better analyze users’ optimal decision from a user perspective as well as design a mechanism from the system designer’s perspective to achieve a desirable outcome.

Oral Session 13: Feature Extraction and Modeling for ASR

Garnet 213-218
10:00 – 12:00, Tuesday 16 September 2014
Chairs: Satoshi Nakamura and Bhuvana Ramabhadran

Acoustic Modeling with Deep Neural Networks Using Raw Time Signal for LVCSR

Zoltán Tüske, Pavel Golik, Ralf Schlüter, Hermann Ney; RWTH Aachen University, Germany

In this paper we investigate how much feature extraction is required by a deep neural network (DNN) based acoustic model for automatic speech recognition (ASR). We decompose the feature extraction pipeline of a state-of-the-art ASR system step by step and evaluate acoustic models trained on standard MFCC features, critical band energies (CRBE), FFT magnitude spectrum and even on the raw time signal. The focus is put on raw time signal as input features, i.e. as much as zero feature extraction prior to DNN training. Noteworthy, the gap in recognition accuracy between MFCC and raw time signal decreases strongly once we switch from sigmoid activation function to rectified linear units, offering a real alternative to standard signal processing. The analysis of the first layer weights reveals that the DNN can discover multiple band pass filters in time domain. Therefore we try to improve the raw time signal based system by initializing the first hidden layer weights with impulse responses of an audiologically motivated filter bank. Inspired by the multi-resolutional analysis layer learned automatically from raw time signal input, we train the DNN on a combination of multiple short-term features. This illustrates how the DNN can learn from the little differences between MFCC, PLP and Gammatone features, suggesting that it is useful to present the DNN with different views on the underlying audio.

Evaluating Robust Features on Deep Neural Networks for Speech Recognition in Noisy and Channel Mismatched Conditions

Vikramjit Mitra, Wen Wang, Horacio Franco, Yun Lei, Chris Bartels, Martin Graciarena; SRI International, USA

Deep Neural Network (DNN) based acoustic models have shown significant improvement over their Gaussian Mixture Model (GMM) counterparts in the last few years. While several studies exist that evaluate the performance of GMM systems under noisy and channel degraded conditions, noise robustness studies on DNN systems have been far fewer. In this work we present a study exploring both conventional DNNs and deep Convolutional Neural Networks (CNN) for noise- and channel-degraded speech recognition tasks using the Aurora4 dataset. We compare the baseline mel-filterbank energies with noise-robust features that we have proposed earlier and show that the use of robust features helps to improve the performance of DNNs or CNNs compared to mel-filterbank energies. We also show that vocal tract length normalization has a positive role in improving the performance of the robust acoustic features. Finally, we show that by combining multiple systems together we can achieve even further improvement in recognition accuracy.

Deep Scattering Spectra with Deep Neural Networks for LVCSR Tasks

Tara N. Sainath 1, Vijayaditya Peddinti 2, Brian Kingsbury 1, Petr Fousek 1, Bhuvana Ramabhadran 1, David Nahamoo 1; 1IBM T.J. Watson Research Center, USA; 2Johns Hopkins University, USA

Log-mel filterbank features, which are commonly used features for CNNs, can remove higher-resolution information from the speech signal. A novel technique, known as Deep Scattering Spectrum (DSS), addresses this issue and looks to preserve this information. DSS features have shown promise on TIMIT, both for classification and recognition. In this paper, we extend the use of DSS features for LVCSR tasks. First, we explore the optimal multi-resolution time and frequency scattering operations for LVCSR tasks. Next, we explore techniques to reduce the dimension of the DSS features. We also incorporate speaker adaptation techniques into the DSS features. Results on a 50 and 430 hour English Broadcast News task show that the DSS features provide between a 4-7% relative improvement in WER over log-mel features, within a state-of-the-art CNN framework which incorporates speaker-adaptation and sequence training. Finally, we show that DSS features are similar to multi-resolution log-mel + MFCCs, and similar improvements can be obtained with this representation.
Robust CNN-Based Speech Recognition with Gabor Filter Kernels
Shuo-Yin Chang, Nelson Morgan; University of California at Berkeley, USA

Tue-O-13-6

As has been extensively shown, acoustic features for speech recognition can be learned from neural networks with multiple hidden layers. However, the learned transformations may not sufficiently generalize to test sets that have a significant mismatch to the training data. Gabor features, on the other hand, are generated from spectro-temporal filters designed to model human auditory processing. In previous work, these features are used as inputs to neural networks, which improved word accuracy for speech recognition in the presence of noise. Here we propose a neural network architecture called a Gabor Convolutional Neural Network (GCNN) that incorporates Gabor functions into convolutional filter kernels. In this architecture, a variety of Gabor features served as the multiple feature maps of the convolutional layer. The filter coefficients are further tuned by back-propagation training. Experiments used two noisy versions of the WSJ corpus: Aurora 4, and RATS re-noised WSJ. In both cases, the proposed architecture performs better than other noise-robust features that we have tried, namely, ETSI-AFE, PNCC, Gabor features without the CNN-based approach, and our best neural network features that don't incorporate Gabor functions.

Probabilistic Linear Discriminant Analysis with Bottleneck Features for Speech Recognition
Liang Lu, Steve Renals; University of Edinburgh, UK

Tue-O-13-5

We have recently proposed a new acoustic model based on probabilistic linear discriminant analysis (PLDA) which enjoys the flexibility of using higher dimensional acoustic features, and is more capable to capture the intra-frame feature correlations. In this paper, we investigate the use of bottleneck features obtained from a deep neural network (DNN) for the PLDA-based acoustic model. Experiments were performed on the Switchboard dataset—a large vocabulary conversational telephone speech corpus. We observe significant word error reduction by using the bottleneck features. In addition, we have also compared the PLDA-based acoustic model to three others using Gaussian mixture models (GMMs), subspace GMMs and hybrid deep neural networks (DNNs), and PLDA can achieve comparable or slightly higher recognition accuracy from our experiments.

Evaluating Speech Features with the Minimal-Pair ABX Task (II): Resistance to Noise
Thomas Schatz 1, Vijayaditya Peddinti 2, Xuan-Nga Cao 1, Francis Bach 3, Hynek Hermansky 2, Emmanuel Dupoux 1; 1LSCP, France; 2Johns Hopkins University, USA; 3INRIA, France

Tue-O-13-3

The Minimal-Pair ABX (MP-ABX) paradigm has been proposed as a method for evaluating speech features for zero-resource/unsupervised speech technologies. We apply it in a phoneme discrimination task on the Articulation Index corpus to evaluate the resistance to noise of various speech features. In Experiment 1, we evaluate the robustness to additive noise at different signal-to-noise ratios, using car and babble noise from the Aurora-4 database and white noise. In Experiment 2, we examine the robustness to different kinds of convolutional noise. In both experiments we consider two classes of techniques to induce noise resistance: smoothing of the time-frequency representation and short-term adaptation in the time-domain. We consider smoothing along the spectral axis (as in PLP) and along the time axis (as in FDLF). For short-term adaptation in the time-domain, we compare the use of a static compressive non-linearity followed by RASTA filtering to an adaptive compression scheme.

Oral Session 14: Speech Analysis I
Peridot 202-203
10:00 – 12:00, Tuesday 16 September 2014
Chairs: Abeer Alwan and Hideki Kawahara

Lateral Formants in Three Central Australian Languages
Marija Tabin 1, Andrew Butterch 2, Gavan Breen 3, Richard Beare 4; 1La Trobe University, Australia; 2Flinders University, Australia; 3Institute for Aboriginal Development, Australia; 4Murdoch Children’s Research Institute, Australia

Tue-O-13-1

This study examines dental, alveolar, retroflex and palatal lateral formants. Data are taken from three languages of Central Australia: Arrerre, Pitjantjatjara and Warlpiri. Results show that in relation to the alveolar lateral, the dental has a lower F1 and a higher F4; the retroflex has lower F3 and F4 and slightly higher F2; and the palatal has lower F1 and higher F2, F3 and F4. These results are discussed in light of various acoustic models of lateral production.

Detecting Articulatory Compensation in Acoustic Data Through Linear Regression Modeling
Alina Khasanova 1, Jennifer Cole 2, Mark Hasegawa-Johnson 2; 1Independent Researcher, Morocco; 2University of Illinois at Urbana-Champaign, USA

Tue-O-13-2

Examining articulatory compensation has been important in understanding how the speech production system is organized, and how it relates to the acoustic and ultimately phonological levels. This paper offers a method that detects articulatory compensation in the acoustic signal, which is based on linear regression modeling of co-variation patterns between acoustic cues. We demonstrate the method on selected acoustic cues for spontaneously produced American English stop consonants. Compensatory patterns of cue variation were observed for voiced stops in some cue pairs, while uniform patterns of cue variation were found for stops as a function of place of articulation or position in the word. Overall, the results suggest that this method can be useful for observing articulatory strategies indirectly from acoustic data and testing hypotheses about the conditions under which articulatory compensation is most likely.

The Relationship Between the Second Subglottal Resonance and Vowel Class, Standing Height, Trunk Length, and F0 Variation for Mandarin Speakers
Jinxin Guo 1, Angli Liu 1, Harish Arskere 1, Abeer Alwan 1, Steven M. Lulich 2; 1University of California at Los Angeles, USA; 2Indiana University, USA

Tue-O-14-3

The relationship between vowel formants and the second subglottal resonance (Sg2) has previously been explored in English, German, Hungarian and Korean. Results from these studies indicate that
vowel space is categorically divided by Sg2 and that Sg2 correlates well with standing height. One of the goals of this work is to verify if the above findings hold true in Mandarin as well. The correlation between Sg2 and sitting height (trunk length) is also studied. Further, since Mandarin is a tonal language (with more pitch variations compared to English), we study the relationship between Sg2 and fundamental frequency (F0). A new corpus of simultaneous recordings of speech and subglottal acoustics was collected from 20 native Mandarin speakers. Results on this corpus indicate that Sg2 divides vowel space in Mandarin as well, and that it is more correlated with sitting height than standing height. Paired t-tests are conducted on the Sg2 measurements from different vowel parts, which represent different F0 regions. Preliminary results show that there is no statistically-significant variation of Sg2 with F0 within a tone.

Comparison of Speech Quality with and without Sensors in Electromagnetic Articulograph AG 501 Recording

Nisha Meenakshi 1, Chiranjeevi Yarra 1, B.K. Yamini 2, Prasanta Kumar Ghosh 1, 1 Indian Institute of Science, India; 2 NIMHANS, India

In the recordings using electromagnetic articulograph AG 501, sensors are glued to subject’s articulators such as jaw, lips and tongue and both speech and articulatory movements are simultaneously recorded. In this work, we study the effect of the presence of the sensors on the quality of speech spoken by the subject. This is done by recording when a subject speaks a set of 19 VCV stimuli while sensors are attached to subject’s articulators. For comparison, we also record the same set of stimuli spoken by the same subject but with no sensors attached to subject’s articulators. Both subjective and objective comparisons are made on the recorded stimuli in these two settings. Subjective evaluation is carried out using 16 evaluators. Listening experiments with recordings from five subjects show that the recordings with sensors attached are significantly different from those without sensors attached in terms of human recognition score as well as on a perceptual difference measure. This is also supported in the objective comparison which computes dissimilarity measure using the spectral shape information.

Impact of Age in the Production of European Portuguese Vowels

Luciana Albuquerque 1, Catarina Oliveira 1, António Teixeira 1, Pedro Sa-Couto 1, João Freitas 1, Miguel Sales Dias 2, 1 Universidade de Aveiro, Portugal; 2 Microsoft, Portugal

The elderly population is quickly increasing in the developed countries. However, in European Portuguese (EP) no studies have examined the impact of age-related structural changes in speech acoustics. The purpose of this paper is to analyse the effect of age (60-70), (71-80) and (81-90), gender and type of vowel in the acoustic characteristics (fundamental frequency (F0), first formant (F1), second formant (F2) and duration) of the EP vowels. A sample of 78 speakers was selected from the database of elderly speech collected by Microsoft Language Development Center (MLDC) within the Living Usability Lab (LUL) project. It was observed that duration is the only parameter that significantly changes with ageing, being the highest value found in the [81-90] group. Moreover, F0 decreases in females and increases in males with ageing. In general, F1 and F2 decreases with ageing, mainly in females. Comparing the data obtained with the results of previous studies with adult speakers, a trend towards the centralization of vowels with ageing is observed. This investigation is the starting point for a broader study which will allow to analyse the changes in vowels acoustics from childhood to old age in EP.

Oral Session 15: Speech Technologies and Applications

Peridot 204-205
10:00 - 12:00, Tuesday 16 September 2014
Chairs: Mark Gales and Dirk van Compernolle

Choosing Useful Word Alternates for Automatic Speech Recognition Correction Interfaces

David Harwath 1, Alexander Gruenstein 2, Ian McGregor 2, 1 MIT, USA; 2 Google, USA

Speech recognition is an increasingly important input modality, especially for mobile computing. Because errors are unavoidable in real applications, efficient correction methods can greatly enhance the user experience. In this paper we study a reranking and classification strategy for choosing word alternates to display to the user in the framework of a tap-to-correct interface. By employing a logistic regression model to estimate the probability that an alternate will offer a useful correction to the user, we can significantly reduce the average length of the alternates lists generated with no reduction in the number of words they are able to correct.

An Initial Investigation of Long-Term Adaptation for Meeting Transcription

X. Chen 1, Mark J.F. Gales 1, Kate M. Knill 1, Catherine Breslin 1, Langzhou Chen 2, K.K. Chin 2, Vincent Wan 2, 1 University of Cambridge, UK; 2 Toshiba Research Europe, UK

Meeting transcription is a very useful and challenging task. The
majority of research to date has focused on individual meeting, or only a small group of meetings. In many practical deployments, multiple related meetings will take place over a long period of time. This paper describes an initial investigation of how this long-term data can be used to improve meeting transcription. A corpus of technical meetings, using a single microphone array, was collected over a two year period, yielding a total of 179 hours of meeting data. Baseline systems using deep neural network acoustic models, in both Tandem and Hybrid configurations, and neural network-based language models are described. The impact of supervised and unsupervised adaptation of the acoustic models is then evaluated, as well as the impact of improved language models.

**Progress in the BBN Keyword Search System for the DARPA RATS Program**

Tim Ng¹, Roger Hsiao¹, Le Zhang¹, Damianos Karakos¹, Sri Harish Mallidi², Martin Karafiát³, Karel Vesely³, Igor Szöke³, Bing Zhang¹, Long Nguyen¹, Richard Schwartz¹; ¹Raytheon BBN Technologies, USA; ²Johns Hopkins University, USA; ³Brno University of Technology, Czech Republic

This paper presents a set of techniques that we used to improve our keyword search system for the third phase of the DARPA RATS (Robust Automatic Transcription of Speech) program, which seeks to advance state of the art detection capabilities on audio from highly degraded radio communication channels. The results for both Levantine and Farsi, which are the two target languages for the keyword search (KWS) task, are reported. About 13% absolute reduction in word error rate (from 70.2% to 57.6%) is achieved by using acoustic features derived from stacked Multi-Layer Perceptrons (MLP) and Deep Neural Network (DNN) acoustic models. In addition to score normalization and score/system combination for keyword search, we showed that the false alarm rate at the target false reject rate (15%) was reduced by about 1% (from 5.39% to 4.45%) by reducing the deletion errors of the speech-to-text system.

**Speech-to-Text Technology to Transcribe and Disclose 100,000+ Hours of Bilingual Documents from Historical Czech and Czechoslovak Radio Archive**

Jan Nouza, Petr Cerva, Jindrich Zdansky, Karel Blavka, Marek Bohac, Jan Silovsky, Josef Chaloupka, Michaela Kucharova, Ladislav Seys, Jiří Malek, Michal Rott; Technical University of Liberec, Czech Republic

In this paper, we present the outcome of a 4-year project whose ultimate goal is to develop a complex platform that can transcribe, index and make searchable the historical archive of Czech and Czechoslovak Radio. The archive covers 90 years of public broadcasting and contains hundreds of thousands of audio documents. The developed modular platform employs our LVCSR system that has to cope with 2 related languages: Czech and Slovak. Furthermore, it must deal with audio files of varying quality (e.g., recordings originally stored on matrices or tapes, data passed through analog and digital telephone lines, speech recorded during parliament or court sessions, etc.). The system includes speaker and language identification modules, a narrow-band signal detector, a music/song detector, and several other components to enhance transcription accuracy and provide support for multi-optional search. We evaluate the performance on broadcast news test sets grouped according to decades. We show that after acoustic and language model adaptation WER values are in range 8-14% and do not differ much since 1960s to present. We report also results achieved on other types of documents (e.g., talk shows, political debates, public speeches, etc), where the WER is higher but still acceptable for most search tasks.

**Automatic Assessment of Children’s Reading with the FLavORe Decoding Using a Phone Confusion Model**

Emre Yılmaz, Joris Peleman, Hugo Van hamme; Katholieke Universiteit Leuven, Belgium

Reading skills of children can be improved with the help of automatic reading tutors (ART), i.e. interactive software with an appealing interface which supports and challenges the child in the reading task, provides instantaneous feedback and automatically assesses its reading skills. For this purpose, ARTs benefit from automatic speech recognition technology for tracking the child’s responses and detecting reading miscues (errors). In previous work, a novel speech recognition architecture has been proposed which adopts a two-layered structure: first a phone recognizer uses task-independent acoustic and language models to generate a phone lattice which is then decoded using a lexicon of expected words and task-dependent finite state grammars. This approach has shown significant improvements in reading miscue detection. In this paper, we extend this technique by employing a more flexible decoding scheme that allows substitution, deletion and insertion of phones. Specifically, the phone lattice generated in the first layer is extended based on a phone confusion matrix that models the typical phone confusions in a language. The proposed system has provided improved miscue detection on the CHORE database compared to a baseline system without a phone confusion model.

**RWTH LVCSR Systems for Quaero and EU-Bridge: German, Polish, Spanish and Portuguese**

M. Ali Basha Shaik, Zoltán Tüske, M. Ali Tahir, Markus Nußbaum-Thom, Ralf Schlüter, Hermann Ney; RWTH Aachen University, Germany

In this paper, German, Polish, Spanish, and Portuguese large vocabulary continuous speech recognition (LVCSR) systems developed by the RWTH Aachen University are presented. All the above mentioned systems for the aforementioned languages are used for the Quaero and EU-Bridge project evaluations. The LVCSR systems developed for these competitive evaluations focus on various domains like broadcast news, podcasts and lecture domain. Transcription of the speech for these tasks is challenging due to huge variability in the acoustic conditions and a significant portion of audio data includes spontaneous speech. Good improvements are obtained using state-of-the-art multilingual bottleneck features, minimum phone error trained acoustic models, language model (LM) adaptation and confusion-network based system combination. In addition, an open vocabulary approach using morphemic units is investigated along with the LM adaptation for the German LVCSR.
Single Channel Source Separation with General Stochastic Networks

Matthias Zöhrer, Franz Pernkopf; Technische Universität Graz, Austria

Single channel source separation (SCSS) is ill-posed and thus challenging. In this paper, we apply general stochastic networks (GSNs) — a deep neural network architecture — to SCSS. We extend GSNs to be capable of predicting a time-frequency representation, i.e. softmask by introducing a hybrid generative-discriminative training objective to the network. We evaluate GSNs on data of the 2nd CHiME speech separation challenge. In particular, we provide results for a speaker dependent, a speaker independent, a matched noise condition and an unmatched noise condition task. We compare to other deep architectures, namely a deep belief network (DBN) and a multi-layer perceptron (MLP). In general, deep architectures perform well on SCSS tasks.

Large-Margin Conditional Random Fields for Single-Microphone Speech Separation

Yu Ting Yeung1, Tan Lee1, Cheung-Chi Leung2; 1Chinese University of Hong Kong, China; 2A*STAR, Singapore

Conditional random field (CRF) formulations for single-microphone speech separation are improved by large-margin parameter estimation. Speech sources are represented by acoustic state sequences from speaker-dependent acoustic models. The large-margin technique improves the classification accuracy of acoustic states by reducing generalization error during the training phase. Non-linear mappings inspired from the mixture-maximization (MIXMAX) model are applied to speech mixture observations. Compared with a factorial hidden Markov model baseline, the improved CRF formulations achieve better separation performance with significantly fewer training data. The separation performance is evaluated in terms of objective speech quality measures and speech recognition accuracy on the reconstructed sources. Compared with the CRF formulations without large-margin parameter estimation, the improved formulations achieve better performance without modifying the statistical inference procedures, especially when the sources are modeled with increased number of acoustic states.

On the Use of the Watson Mixture Model for Clustering-Based Under-Determined Blind Source Separation

Ingrid Jafari1, Roberto Togneri1, Sven Nordholm2; 1University of Western Australia, Australia; 2Curtin University of Technology, Australia

In this paper, we investigate the application of a generative clustering technique for the estimation of time-frequency source separation masks. Recent advances in time-frequency clustering-based approaches to blind source separation have touched upon the Watson mixture model (WMM) as a tool for source separation. However, most methods have been frequency bin-wise and have thus required the additional permutation alignment stage, and previous full-band methods which employ the WMM have yet to be applied to the under-determined setting. We propose to evaluate the clustering ability of the WMM within the clustering-based source separation framework. Evaluations confirm the superiority of the WMM against other previously used clustering techniques such as the fuzzy c-means.

Binary Mask Estimation Based on Frequency Modulations

Chung-Chien Hsu, Jen-Tzung Chien, Tai-Shih Chi; National Chiao Tung University, Taiwan

In this paper, a binary mask estimation algorithm is proposed based on modulations of speech. A multi-resolution spectro-temporal analytical auditory model is utilized to extract modulation features to estimate the binary mask, which is often used in speech segregation applications. The proposed method estimates noise from the beginning of each test sentence, a common approach seen in many conventional speech enhancement algorithms, to further enhance the modulation features. Experimental results demonstrate that the proposed method outperforms the AMS-GMM system in terms of the HIT-FA rate when estimating the binary mask.

Bayesian Factorization and Selection for Speech and Music Separation

Po-Kai Yang, Chung-Chien Hsu, Jen-Tzung Chien; National Chiao Tung University, Taiwan

This paper proposes a new Bayesian nonnegative matrix factorization (NMF) for speech and music separation. We introduce the Poisson likelihood for NMF approximation and the exponential prior distributions for the factorized basis matrix and weight matrix. A variational Bayesian (VB) EM algorithm is developed to implement an efficient solution to variational parameters and model parameters for Bayesian NMF. Importantly, the exponential prior parameter is used to control the sparseness in basis representation. The variational lower bound in VB-EM procedure is derived as an objective to conduct adaptive basis selection for different mixed signals. The experiments on single-channel speech/music separation show that the adaptive basis representation in Bayesian NMF via model selection performs better than the NMF with the fixed number of bases in terms of signal-to-distortion ratio.

Self-Adaption in Single-Channel Source Separation

Michael Wohlmayr1, Ludwig Mohr2, Franz Pernkopf2; 1Commend International, Austria; 2Technische Universität Graz, Austria

Single-channel source separation (SCSS) usually uses pre-trained source-specific models to separate the sources. These models capture the characteristics of each source and they perform well when matching the test conditions.

In this paper, we extend the applicability of SCSS. We develop an EM-like iterative adaption algorithm which is capable to adapt the pre-trained models to the changed characteristics of the specific situation, such as a different acoustic channel introduced by variation in the room acoustics or changed speaker position. The adaption framework requires signal mixtures only, i.e. specific single source signals are not necessary. We consider speech/noise mixtures and we restrict the adaption to the speech model only.
Voice command system in multi-room smart homes for assisting people in loss of autonomy in their daily activities faces several challenges, one of which being the distant condition which impacts the ASR system performance. This paper presents an approach to improve voice command recognition at the decoding level by using multiple sources and model adaptation. The method has been tested on data recorded with 11 elderly and visually impaired participants in a real smart home. The results show an error rate of 3.2% in off-line condition and of 13.2% in on-line condition.

In this paper, we investigate unsupervised acoustic model training for dysarthric speech. These models are first, frame-based Gaussian posteriorgrams, obtained from Vector Quantization (VQ), second, so-called Acoustic Unit Descriptors (AUDs), which are hidden Markov models of phone-like units, that are trained in an unsupervised fashion, and, third, posteriorgrams computed on the AUDs. Experiments were carried out on a database collected from a home automation task and containing three people scenarios, respectively. The recognition accuracy and manner of articulation tracking accuracy (MOTA) was 62.81% and 72.31% for single person multiple objects tracking precision (MOTP) of the developed video monitoring system was 0.78 and 0.73 and the multiple objects tracking precision (MOTP) of the developed video monitoring system was 0.78 and 0.73 and the multiple objects tracking precision (MOTP). The study presented in this paper extends previous work by investigating to what extent information about the speech production process can be recovered from PMA data. In particular, the three main aspects of speech production are investigated here: voicing, place of articulation and manner of articulation. The results show that PMA achieves comparable accuracy to using the audio signal for discriminating the place of articulation of consonants, but it provides little information regarding the voicing and manner of articulation.

In the assisted living environment, users are tracked by some omnidirectional video cameras, as well as speech and non-speech audio events are recognized by an array of microphones. The multiple objects tracking precision (MOTA) of the developed video monitoring system was 62.81% and 72.31% for single person and three people scenarios, respectively. The recognition accuracy of the proposed multilingual speech and audio events recognition system was 96.3% and 93.8% for user's speech commands and non-speech acoustic events, correspondingly. The design of the assisted living environment, the certain test scenarios and the process of audio-visual database collection are described in the paper.

Analysis of Phonetic Similarity in a Silent Speech Interface Based on Permanent Magnetic Articulography
Jose A. Gonzalez 1, Lam A. Cheah 2, Jie Bai 2, Stephen R. Ell 3, James M. Gilbert 2, Roger K. Moore 1, Phil D. Green 1; 1University of Sheffield, UK; 2University of Hull, UK; 3Castle Hill Hospital, UK

This paper investigates the potential of a silent speech interface (SSI) based on Permanent Magnetic Articulography (PMA) to be used in applications involving unconstrained, phonetically rich speech. In previous work the SSI was evaluated on isolated-word and connected-digits recognition tasks with promising results. Furthermore, it was shown that PMA data is enough to distinguish between minimal pairs of consonants with the same manner and place of articulation but different voicing. The study presented in this paper extends previous work by investigating to what extent information about the speech production process can be recovered from PMA data. In particular, the three main aspects of speech production are investigated here: voicing, place of articulation and manner of articulation. The results show that PMA achieves comparable accuracy to using the audio signal for discriminating the place of articulation of consonants, but it provides little information regarding the voicing and manner of articulation.

Audio-Visual Signal Processing in a Multimodal Assisted Living Environment
Alexey Karpov 1, Lale Akarun 2, Hülya Yalçın 3, Alexander Ronzhin 1, Barış Evrim Demiröz 2, Aysun Çoban 2, Miloš Železný 4; 1Russian Academy of Sciences, Russia; 2Boğaziçi Üniversitesi, Turkey; 3İstanbul Teknik Üniversitesi, Turkey; 4University of West Bohemia, Czech Republic

In this paper, we present some novel methods and applications for audio and video signal processing for a multimodal environment of an assisted living smart space. This intelligent environment was developed during the 7th Summer Workshop on Multimodal Interfaces eNTERFACE. It integrates automatic systems for audio and video-based monitoring and user tracking in the smart space. In the assisted living environment, users are tracked by some omnidirectional video cameras, as well as speech and non-speech audio events are recognized by an array of microphones. The multiple objects tracking precision (MOTA) of the developed video monitoring system was 0.78 and 0.73 and the multiple objects tracking accuracy (MOTA) was 62.81% and 72.31% for single person and three people scenarios, respectively. The recognition accuracy of the proposed multilingual speech and audio events recognition system was 96.3% and 93.8% for user’s speech commands and non-speech acoustic events, correspondingly. The design of the assisted living environment, the certain test scenarios and the process of audio-visual database collection are described in the paper.

On the Selection of the Impulse Responses for Distant-Speech Recognition Based on Contaminated Speech Training
Mirco Ravanelli, Maurizio Omologo; FBK, Italy

Distant-speech recognition represents a technology of fundamental importance for future development of assistive applications.
characterized by flexible and unobtrusive interaction in home environments. State-of-the-art speech recognition still exhibits lack of robustness, and an unacceptable performance variability, due to environmental noise, reverberation effects, and speaker position. In the past, multi-condition training and contamination methods were explored to reduce the mismatch between training and test conditions. However, the performance evaluation can be biased by factors as limited number of positions of speaker and microphones, adopted set of impulse responses, vocabulary and grammars defining the recognition task. The purpose of this paper is to investigate in more detail some critical aspects that characterize such experimental context. To this purpose, our work addressed a microphone network distributed over different rooms of an apartment and a related set of speaker-microphone pairs leading to a very large set of impulse responses. Besides simulations, the experiments also tackled real speech interactions. The performance evaluation was based on a phone-loop task, in order to minimize the influence of linguistic constraints. The experimental results show how less critical is an accurate selection of impulse responses, if compared to other factors as the signal-to-noise ratio introduced by additive background noise.

Adaptive Speech Recognition and Dialogue Management for Users with Speech Disorders
I. Casanueva, H. Christensen, Thomas Hain, Phil D. Green; University of Sheffield, UK

Spoken control interfaces are very attractive to people with severe physical disabilities who often also have a type of speech disorder known as dysarthria. This condition is known to decrease the accuracy of automatic speech recognisers (ASRs) especially for users with moderate to severe dysarthria. In this paper we investigate how applying probabilistic dialogue management (DM) techniques can improve interaction performance of an environmental control system for such users. The effect of having access to different amounts of adaptation data, as well as using different vocabulary size for speakers of different intelligibilities is investigated. We explore the effect of adapting the DM models as the ASR performance increases, such as is the case in systems where more adaptation data is collected through system use. Improvements compared to a non-probabilistic DM baseline are seen both in terms of dialogue length and success rate, 9% and 25% mean relative improvement respectively. Looking at just the more severe dysarthric speakers these numbers rise 25% and 75% mean relative improvement. These improvements are higher when the ASR data adaptation amount is small. Further results show that a DM trained on data from multiple speakers outperform a DM trained on data from a single speaker.

Prediction of Cognitive Performance in an Animal Fluency Task Based on Rate and Articulatory Markers
Bea Yu 1, Thomas F. Quattori 1, James R. Williamson 1, James C. Mundt 2; 1 MIT Lincoln Laboratory, USA; 2 Center for Telepsychology, USA

Neuropsychological changes in the brain associated with early dementia can disrupt articulatory timing and precision in speech production. Motivated by this observation, we address the hypothesis that speaking rate and articulatory coordination, as manifested through formant frequency tracks, can predict performance on an animal fluency task administered to the elderly. Specifically, using phoneme-based measures of speaking rate and articulatory coordination derived from formant cross-correlation measures, we investigate the capability of speech features, estimated from paragraph-recall and naturalistic free speech, to predict animal fluency assessment scores. Using a database consisting of audio from elderly subjects over a 4 year period, we develop least-squares regression models of our cognitive performance measures. The best performing model combined speaking rate and formant features, resulting in a correlation (R) of 0.61 and a root mean squared error (RMSE) of 3.07 with respect to a 9-34 score range. Vocal features thus provide a reduction by about 30% in MSE from a baseline (mean score) in predicting cognitive performance derived from the animal fluency assessment.

Analysis of Laughter Events in Real Science Classes by Using Multiple Environment Sensor Data
Carlos Ishi, Hiroaki Hatano, Norihiro Hagita; ATR IRC, Japan

The extraction of sound events in environments where a large number of people are present is a challenging problem. In order to tackle that problem, we have been developing a sound environment intelligence system which is able to get information about who is talking, where and when, based on integration of multiple microphone arrays and human tracking technologies. We installed the developed system in a science room of an elementary school, and collected data of real science classes during a period of one month. In the present paper, among the sound activities appearing in the science classes, we focused on the analysis of laughter events, considering that laughter conveys important social functions in communication. Laughter events were extracted by making use of visual displays of spatial-temporal information provided by the developed system. Subjective evaluation of the laughter events revealed relationship between the laughter type (including production, style, and vowel-quality aspects), the functions in communication, and the appropriateness in the classroom context.

Poster Session 10: DNN for ASR
Max Atria Gallery
10:00 - 12:00, Tuesday 16 September 2014
Chair: Anton Ragni

Parallel Deep Neural Network Training for LVCSR Tasks Using Blue Gene/Q
Tara N. Sainath, I-hsin Chung, Bhuvana Ramabhadran, Michael Picheny, John Gunnels, Brian Kingsbury, George Saon, Vernon Austel, Upendra Chaudhari; IBM T.J. Watson Research Center, USA

While Deep Neural Networks (DNNs) have achieved tremendous success for LVCSR tasks, training these networks is slow. To date, the most common approach to train DNNs is via stochastic gradient descent (SGD), serially on a single CPU machine. Serial training, coupled with the large number of training parameters and speech data set sizes, makes DNN training very slow for LVCSR tasks. While 2nd order, data-parallel methods have also been explored, these methods are not always faster on CPU clusters due to the large communication cost between processors. In this work, we explore using a specialized hardware/software approach, utilizing a Blue Gene/Q (BG/Q) system, which has thousands of processors and excellent inter-processor communication. We explore using the 2nd order Hessian-free (HF) algorithm for DNN training with BG/Q, for both cross-entropy and sequence training of DNNs. Results on three LVCSR tasks indicate that using HF with BG/Q
offers up to an 11x speedup, as well as an improved word error rate (WER), compared to SGD on a GPU.

Word Embeddings for Speech Recognition
Samy Bengio, Georg Heigold; Google, USA
Tue-P-10-2, Poster

Speech recognition systems have used the concept of states as a way to decompose words into sub-word units for decades. As the number of such states now reaches the number of words used to train acoustic models, it is interesting to consider approaches that relax the assumption that words are made of states. We present here an alternative construction, where words are projected into a continuous embedding space where words that sound alike are nearby in the Euclidean sense. We show how embeddings can still allow to score words that were not in the training dictionary. Initial experiments using a lattice rescoring approach and model combination on a large realistic dataset show improvements in word error rate.

1-Bit Stochastic Gradient Descent and its Application to Data-Parallel Distributed Training of Speech DNNs
Frank Seide 1, Hao Fu 1, Jasha Droppo 2, Gang Li 1, Dong Yu 2; 1Microsoft, China; 2Microsoft, USA
Tue-P-10-3, Poster

We show empirically that in SGD training of deep neural networks, one can, at no or nearly no loss of accuracy, quantize the gradients aggressively — to but one bit per value — if the quantization error is carried forward across minibatches (error feedback). This size reduction makes it feasible to parallelize SGD through data-parallelism with fast processors like recent GPUs. We implement data-parallel deterministically distributed SGD by combining this finding with AdaGrad, automatic minibatch-size selection, double buffering, and model parallelism. Unexpectedly, quantization benefits AdaGrad, giving a small accuracy gain.

For a typical Switchboard DNN with 46M parameters, we reach 68% efficiency.

Boundary Contraction Training for Acoustic Models Based on Discrete Deep Neural Networks
Ryu Takeda, Naoyuki Kanda, Nobuo Nukaga; Hitachi, Japan
Tue-P-10-4, Poster

The boundary contraction training for acoustic models based on deep neural networks in a discrete system (discrete DNNs) is presented in this paper. Representing the parameters of DNNs with small bits (such as 4 bits) can reduce not only memory usage but also computational complexity by utilizing a CPU cache and look-up tables efficiently. However, simply quantizing parameters of the normal continuous DNNs degrades the recognition accuracy seriously. We tackle this problem by developing a specialized training algorithm for discrete DNNs. In our algorithm, continuous DNNs with boundary constraint are first trained, and the trained parameters are then quantized to meet the representation of discrete DNNs. When training continuous DNNs, we introduce the boundary contraction mapping to shrink the distribution of parameters for reducing the quantization error. In our experiments with 4-bit discrete DNNs, while simply quantizing normally trained DNNs degrades the word accuracy by more than 50 points, our method can maintain the high word accuracy of DNNs with only two points degradation.

Restructuring Output Layers of Deep Neural Networks Using Minimum Risk Parameter Clustering
Yotaro Kubo 1, Jun Suzuki 2, Takaaki Hori 2, Atsushi Nakamura 1, 3; 1Amazon, Germany; 2NTT Corporation, Japan; 3Nagoya City University, Japan
Tue-P-10-5, Poster

This paper attempts to optimize a topology of hidden Markov models (HMMs) for automatic speech recognition. Current state-of-the-art acoustic models for ASR involve HMMs with deep neural network (DNN)-based emission density functions. Even though DNN parameters are typically trained by optimizing a discriminative criterion, topology optimization of HMMs is usually performed by optimizing a generative criterion. Several approaches have been studied to achieve a discriminative state clustering, these approaches typically assume underlying Gaussian distributions of the acoustic features, and do not compatible with DNN-based emission density functions. In this paper, we attempt to derive a discriminative restructuring method of an HMM topology by introducing discriminative optimization with discrete constraints on the parameters, which force the parameters to be tied with the parameters of the other states. By applying this constrained optimization to the clustering of parameters of DNN-based acoustic models, we derived a discriminative HMM restructuring method that maintains discriminative performance of the original HMMs with the large number of states.

Distributed Asynchronous Optimization of Convolutional Neural Networks
William Chan, Ian Lane; Carnegie Mellon University, USA
Tue-P-10-6, Poster

Recently, deep Convolutional Neural Networks have been shown to outperform Deep Neural Networks for acoustic modelling, producing state-of-the-art accuracy in speech recognition tasks. Convolutional models provide increased model robustness through the usage of pooling invariance and weight sharing across spectrum and time. However, training convolutional models is a very computationally expensive optimization procedure, especially when combined with large training corpora. In this paper, we present a novel algorithm for scalable training of deep Convolutional Neural Networks across multiple GPUs. Our distributed asynchronous stochastic gradient descent algorithm incorporates sparse gradients, momentum and gradient decay to accelerate the training of these networks. Our approach is stable, neither requiring warm-starting or excessively large minibatches. Our proposed approach enables convolutional models to be efficiently trained across multiple GPUs, enabling a model to be scaled asynchronously across 5 GPU workers with 40% efficiency.

Convolutional Deep Maxout Networks for Phone Recognition
László Tóth; MTA-SZTE RGAI, Hungary
Tue-P-10-7, Poster

Convolutional neural networks have recently been shown to outperform fully connected deep neural networks on several speech recognition tasks. Their superior performance is due to their
convolutional structure that processes several, slightly shifted versions of the input window using the same weights, and then pools the resulting neural activations. This pooling operation makes the network less sensitive to translations. The convolutional network results published up till now used sigmoid or rectified linear neurons. However, quite recently a new type of activation function called the maxout activation has been proposed. Its operation is closely related to convolutional networks, as it applies a similar pooling step, but over different neurons evaluated on the same input. Here, we conducted experiments with deep convolutional neural networks built from maxout neurons. Phone recognition tests on the TIMIT database show that switching to maxout units from rectifier units decreases the phone error rate for each network configuration studied, and yields relative error rate reductions of between 2% and 6%.

**Joint Sequence Training of Phone and Grapheme Acoustic Model Based on Multi-Task Learning Deep Neural Networks**

Dongpeng Chen 1, Brian Mak 1, Sunil Sivadas 2; 1 HKUST, China; 2 A*STAR, Singapore

Multi-task learning (MTL) can be an effective way to improve the generalization performance of singly learning tasks if the tasks are related, especially when the amount of training data is small. Our previous work applied MTL to the joint training of triphone and trigrapheme acoustic models using deep neural networks (DNNs) for low-resource speech recognition. Significant recognition improvement over the performance of their DNNs trained by single-task learning (STL) was obtained. In that work, both STL-DNNs and MTL-DNNs were trained by minimizing the total frame-wise cross entropies. Since phoneme and grapheme recognition are inherently sequence classification tasks, here we study the effect of sequence-discriminative training on their joint estimation using MTL-DNNs. Experimental evaluation on TIMIT phoneme recognition shows that joint sequence training outperforms frame-wise training of phone and grapheme MTL-DNNs significantly.

**Improving Semi-Supervised Deep Neural Network for Keyword Search in Low Resource Languages**

Roger Hsiao, Tim Ng, Le Zhang, Shivesh Ranjan, Stavros Tsakalidis, Long Nguyen, Richard Schwartz; Raytheon BBN Technologies, USA

In this work, we investigate how to improve semi-supervised DNN for low resource languages where the initial systems may have high error rate. We propose using semi-supervised MLP features for DNN training, and we also explore using confidence to improve semi-supervised cross entropy and sequence training. The work conducted in this paper was evaluated under the IARPA Babel multi-condition training methods providing significant gains in levels are used. Experimental results show the first and third levels are used. Lastly, a four-condition training setup is proposed where models for clean, mild, moderate, and severe reverberation are used. The second utilizes two-condition training where clean and reverberant models are used. Lastly, a four-condition training setup is proposed where models for clean, mild, moderate, and severe reverberation levels are used. Experimental results show the first and third multi-condition training methods providing significant gains in performance relative to the baseline, with the latter being more suitable for practical resource-constrained far-field applications.

**Pruning Deep Neural Networks by Optimal Brain Damage**

Chao Liu, Zhiyong Zhang, Dong Wang; Tsinghua University, China

A main advantage of the deep neural network (DNN) model lies on the fact that no artificial assumptions are placed on the data distribution and model structure, which offers the possibility to learn very flexible models. This flexibility, however, may lead to highly redundant parameters, hence demanding computation and risk of over-fitting. Network pruning cuts off unimportant connections, and therefore can be used to produce parsimonious and well generalizable models.

This paper proposes to utilize optimal brain damage (OBD) to conduct DNN pruning. OBD computes connection salience based on Hessians, and thus is sound in theory and reliable in practice. We present our implementation of OBD for DNNs, and demonstrate that the OBD pruning can produce very sparse DNNs while retaining the discriminative power of the original network to a large extent. By comparing with a simple magnitude-based pruning, we find that for weak pruned networks, pruning methods are unimportant since retraining can largely recover the function loss caused by pruning; while for highly pruned networks, sophisticated pruning methods (such as OBD) are clearly superior.

**Poster Session 11: Speaker Recognition — General Topics**

Max Atria Gallery
10:00 - 12:00, Tuesday 16 September 2014
Chair: Nicholas Evans

**Improving the Performance of Far-Field Speaker Verification Using Multi-Condition Training: The Case of GMM-UBM and i-Vector Systems**

Anderson R. Avila 1, Milton Sarria-Paja 1, Francisco J. Fraña 2, Douglas O’Shaughnessy 1, Tiago H. Falk 1; 1 INRS-EMT, Canada; 2 UFABC, Brazil

While considerable work has been done to characterize the detrimental effects of channel variability on automatic speaker verification (ASV) performance, little attention has been paid to the effects of room reverberation. This paper investigates the effects of room acoustics on the performance of two far-field ASV systems: GMM-UBM (Gaussian mixture model - universal background model) and i-vector. We show that ASV performance is severely affected by reverberation, particularly for i-vector based systems. Three multi-condition training methods are then investigated to mitigate such detrimental effects. The first uses matched train/test speaker models based on estimated reverberation time (RT) values. The second utilizes two-condition training where clean and reverberant models are used. Lastly, a four-condition training setup is proposed where models for clean, mild, moderate, and severe reverberation levels are used. Experimental results show the first and third multi-condition training methods providing significant gains in performance relative to the baseline, with the latter being more suitable for practical resource-constrained far-field applications.

**Clustering-Based I-Vector Formulation for Speaker Recognition**

Hung-Shin Lee 1, Yu Tsao 2, Hsin-Min Wang 2, Shyh-Kang Jeng 1; 1 National Taiwan University, Taiwan; 2 Academia Sinica, Taiwan

In this paper, we first reformulate the derivation of the conventional i-vector scheme, which is the state-of-the-art utterance representation for speaker verification, as a modeling of universal background model (UBM)-based mixtures of factor analyzers (UMFA), and then propose a clustering-based UMFA method called
CMFA. In UMFA, each analyzer is characterized by a subspace, and the same projection coordinate of an utterance into individual subspaces is called the i-vector. We relax this assumption by grouping the mixture components of the UBM into clusters according to their acoustic traits. Therefore, in CMFA, each utterance is represented by multiple i-vectors, each of which generated by similar subspaces associated with a same cluster. We also investigate two strategies for merging these i-vectors into a single one to be applied in the classifier of the conventional i-vector framework. The results of experiments conducted on the male portion of the core task in the NIST 2005 Speaker Recognition Evaluation (SRE) in terms of normalized decision cost function (minDCF) and equal error rate (EER) demonstrate the merits of the new i-vector method over the conventional i-vector method.

Speaker Recognition via Fusion of Subglottal Features and MFCCs
Harish Arskire, Hitesh Anand Gupta, Abeer Alwan; University of California at Los Angeles, USA

Motivated by the speaker-specificity and stationarity of subglottal acoustics, this paper investigates the utility of subglottal cepstral coefficients (SGCCs) for speaker identification (SID) and verification (SV). SGCCs can be computed using accelerometer recordings of subglottal acoustics, but such an approach is infeasible in real-world scenarios. To estimate SGCCs from speech signals, we adopt the Bayesian minimum mean squared error (MMSE) estimator proposed in the speech-to-articulatory inversion literature. The joint distribution of SGCCs and speech MFCCs is modeled using the WashU-UCLA corpus (containing simultaneous recordings of speech and subglottal acoustics), and the resulting model is used to obtain an MMSE estimate of SGCCs from unseen (test) MFCCs. Cross-validation experiments on the WashU-UCLA corpus show that the estimation efficacy, on average, is speaker dependent. A score-level fusion of MFCC and SGCC systems outperforms the MFCC-only baseline in both SID and SV tasks. On the TIMIT database (SID), the relative reduction in identification error is 16, 40 and 51% for G.712-filtered (300–3400 Hz), narrowband (0–4000 Hz) and wideband (0–8000 Hz) speech, respectively. On the NIST 2008 database (SV), the relative reduction in equal error rate is 4 and 11% for 10 and 5 second utterances, respectively.

The NIST SRE Summed Channel Speaker Recognition System
Hanwu Sun, Bin Ma; A*STAR, Singapore

This paper presents an improved speaker recognition system for the summed channel evaluation tasks in the 2008 NIST SRE (SRE08) with multiple summed-channel excerpts for speaker training and one summed-channel excerpt for testing. The system includes three main modules in which a hybrid speaker purification and clustering algorithm is adopted to segregate the summed-channel speech, a common speaker identification is proposed by mapping multiple summed-channel excerpts for a common speaker cluster, and the GMM-SVM-NAP algorithm is used for the speaker recognition system. The system achieves an overall EER of 7.82% for all the trials and 4.19% for English trials in the SRE08 3summed-summed task.

Advantages of Wideband Over Narrowband Channels for Speaker Verification Employing MFCCs and LFCCs
Laura Fernández Gallardo1, Michael Wagner1, Sebastian Möller2; 1University of Canberra, Australia; 2T-Labs, Germany

Wideband communications permit the transmission of an extended frequency range compared to the traditional narrowband. While benefits for automatic speaker recognition can be expected, the extent of the contribution of the additional bandwidth in wideband is still unclear. This work compares the i-vector speaker verification performance with the extended bandwidth, and also that the linear scale can lead to better results for narrowband signals. The advantages of linear- over mel-scaled features for wideband depend on the speakers’ gender and on the channel distortion.

Speaker Verification and Spoken Language Identification Using a Generalized I-Vector Framework with Phonetic Tokenizations and Tandem Features
Ming Li, Wenbo Liu; Sun Yat-sen University, China

This paper presents a generalized i-vector framework with phonetic tokenizations and tandem features for speaker verification as well as language identification. First, the tokens for calculating the zero-order statistics is extended from the MFCC trained Gaussian Mixture Models (GMM) components to phonetic phonemes, 3-grams and tandem feature trained GMM components using phoneme posterior probabilities. Second, given the calculated zero-order statistics (posterior probabilities on tokens), the feature used to calculate the first-order statistics is also extended from MFCC to tandem features and is not necessarily the same feature employed by the tokenizer. Third, the zero-order and first-order statistics vectors are then concatenated and represented by the simplified supervised i-vector approach followed by the standard back end modeling methods. We study different system setups with different tokens and features. Finally, selected effective systems are fused at the score level to further improve the performance. Experimental results are reported on the NIST SRE 2010 common condition 5 female part task and the NIST LRE 2007 closed set 30 seconds task for speaker verification and language identification, respectively. The proposed generalized i-vector framework outperforms the i-vector baseline by relatively 45% in terms of equal error rate (EER) and norm minDCF values.

Feature Switching in the i-Vector Framework for Speaker Verification
Asha T.; 1, Saranya M.S.1, Karthik Pandia D.S.1, Srikant Madikeri2, Hema A. Murthy1; 1IIT Madras, India; 2EPFL, Switzerland

Feature fusion is a paradigm that has found success in a number of speech related tasks. The primary objective in applying fusion is to leverage the complementary information present in the features.
Conventionally, either early or late fusion is employed. Early fusion leads to large dimensional feature vectors. Further, the range of feature values for different streams require appropriate normalization. Late fusion is carried out at score level, where the contribution from each type of feature is determined from the set of weights used. Feature switching is yet another paradigm that attempts to capture the diversity in the feature types used. Feature switching gains significance particularly in the context of speaker verification, where the feature type that best discriminates a speaker is used to verify the claims corresponding to that speaker. Earlier, feature switching was attempted in the conventional UBM-GMM framework. In this paper, the idea is extended to the Total Variability Space (TVS) framework. Two different feature types namely Modified Group Delay (MGD) and Mel-Frequency Cepstral Coefficients (MFCC) are explored in the proposed framework. Results are presented on NIST 2010 male database for the speaker verification task.

**PLDA Modeling in the Fishervoice Subspace for Speaker Verification**  
**Jinhua Zhong**, Weiwu Jiang, Wei Rao, Man-Wai Mak, Helen Meng, Chinese University of Hong Kong, China; Hong Kong Polytechnic University, China  
**Poster**  
We have previously developed a Fishervoice framework that maps the JFA-mean supervectors into a compressed discriminant subspace using nonparametric Fishers discriminant analysis. It was shown that performing cosine distance scoring (CDS) on these Fishervoice projected vectors (denoted as f-vectors) can outperform the classical joint factor analysis. Unlike the i-vector framework in which channel variability is suppressed in the classification stage, in the Fishervoice framework, channel variability is suppressed when the f-vectors are constructed. In this paper, we investigate whether channel variability can be further suppressed by performing Gaussian probabilistic discriminant analysis (PLDA) in the classification stage. We also use random subspace sampling to enrich the speaker discriminative information in the f-vectors. Experiments on NIST SRE10 show that PLDA can boost the performance of Fishervoice in speaker verification significantly by a relative decrease of 14.4% in minDCF (from 0.526 to 0.450).

**Performance Factor Analysis for the 2012 NIST Speaker Recognition Evaluation**  
**Alvin F. Martin**, Craig S. Greenberg, Vincent M. Stanford, John M. Howard, George R. Doddington, John J. Godfrey, Systems Plus, USA; NIST, USA; Systems Plus, USA; Johns Hopkins University, USA  
**Poster**  
The 2012 NIST Speaker Recognition Evaluation, held in the autumn of 2012, was designed to examine a variety of factors affecting the performance of automatic systems for speaker recognition. Here we examine, for leading systems included in this evaluation, the observed effects on performance of five such factors: the inclusion in test segment speech of environmental noise or of added synthetic noise of one of three types and one of two intensity levels, the duration of test segment speech, the number and the channel type of target speaker training sessions, the type of the microphone channel used in test segment speech, and the sex of the target speaker. This evaluation is notable for being the first in the series to include examination of the effects on performance of synthetic added noise. The greater impact of crowd noise compared to HVAC noise, and of single speaker noise compared to crowd noise is observed. Future evaluation plans are also discussed.

**Poster Session 12: Speech Processing with Multi-modalities**  
**Max Atria Gallery**  
10:00 - 12:00, Tuesday 16 September 2014  
Chair: Hsin-Min Wang

**Dynamic Stream Weight Estimation in Coupled-HMM-Based Audio-Visual Speech Recognition Using Multilayer Perceptrons**  
**Ahmed Hussen Abdelaziz**, Dorothea Kolossa; Ruhr-Universität Bochum, Germany  
**Poster**  
Jointly using audio and video features can increase the robustness of automatic speech recognition systems in noisy environments. A systematic and reliable performance gain, however, is only achieved if the contributions of the audio and video stream to the decoding decision are dynamically optimized, for example via so-called stream weights. In this paper, we address the problem of dynamic stream weight estimation for coupled-HMM-based audio-visual speech recognition. We investigate the multilayer perceptron (MLP) for mapping reliability measure features to stream weights. As an input for the multilayer perceptron, we use a feature vector containing different model-based and signal-based reliability measures. Training of the multilayer perceptron has been achieved using dynamic oracle stream weights as target outputs, which are found using a recently proposed expectation maximization algorithm. This new approach of MLP-based stream-weight estimation has been evaluated using the Grid audio-visual corpus and has outperformed the best baseline performance, yielding a 23.72% average relative error rate reduction.

**Lipreading using Convolutional Neural Network**  
**Kuniaki Noda**, Yuki Yamauchi, Kazuhiro Nakadai, Hiroshi G. Okuno, Tetsuya Ogata; Waseda University, Japan; Kyoto University, Japan; Honda Research Institute Japan, Japan  
**Poster**  
In recent automatic speech recognition studies, deep learning architecture applications for acoustic modeling have eclipsed
conventional sound features such as Mel-frequency cepstral coefficients. However, for visual speech recognition (VSR) studies, handcrafted visual feature extraction mechanisms are still widely utilized. In this paper, we propose to apply a convolutional neural network (CNN) as a visual feature extraction mechanism for VSR. By training a CNN with images of a speaker’s mouth area in combination with phoneme labels, the CNN acquires multiple convolutional filters, used to extract visual features essential for recognizing phonemes. Further, by modeling the temporal dependencies of the generated phoneme label sequences, a hidden Markov model in our proposed system recognizes multiple isolated words. Our proposed system is evaluated on an audio-visual speech dataset comprising 300 Japanese words with six different speakers. The evaluation results of our isolated word recognition experiment demonstrate that the visual features acquired by the CNN significantly outperform those acquired by conventional dimensionality compression approaches, including principal component analysis.

Lipreading Approach for Isolated Digits Recognition Under Whisper and Neutral Speech

Fei Tao, Carlos Busso; University of Texas at Dallas, USA

Whisper is a speech production mode normally used to protect confidential information. Given the differences in the acoustic domain, the performance of automatic speech recognition (ASR) systems decreases with whisper speech. An appealing approach to improve the performance is the use of lipreading. This study explores the use of visual features characterizing the lips’ geometry and appearance to recognize digits under normal and whisper speech conditions using hidden Markov models (HMMs). We evaluate the proposed features on the digit part of the audiovisual whisper (AVW) corpus. While the proposed system achieves high accuracy in speaker dependent conditions (80.8%), the performance decreases when we evaluate speaker independent models (52.9%). We propose supervised adaptation schemes to reduce the mismatch between speakers. Across all conditions, the performance of the classifiers remain competitive even in the presence of whisper speech, highlighting the benefits of using visual features.

Multimodal Exemplar-Based Voice Conversion Using Lip Features in Noisy Environments

Kenta Masaka, Ryo Aihara, Tetsuya Takiguchi, Yasuo Arik; Kobe University, Japan

This paper presents a multimodal voice conversion (VC) method for noisy environments. In our previous exemplar-based VC method, source exemplars and target exemplars are extracted from parallel training data, in which the same texts are uttered by the source and target speakers. The input source signal is then decomposed into source exemplars, noise exemplars obtained from the input signal, and their weights. Then, the converted speech is constructed from the target exemplars and the weights related to the source exemplars. In this paper, we propose a multimodal VC method that improves the noise robustness of our previous exemplar-based VC method. As visual features, we use not only conventional DCT but also the features extracted from Active Appearance Model (AAM) applied to the lip area of a face image. Furthermore, we introduce the combination weight between audio and visual features and formulate a new cost function in order to estimate the audio-visual exemplars. By using the joint audio-visual features as source features, the VC performance is improved compared to a previous audio-input exemplar-based VC method. The effectiveness of this method was confirmed by comparing its effectiveness with that of a conventional Gaussian Mixture Model (GMM)-based method.

Towards a Practical Silent Speech Recognition System

Yunbin Deng1, James T. Heaton2, Geoffrey S. Meltzner1; 1BAE Systems, USA; 2Massachusetts General Hospital, USA

Our recent efforts towards developing a practical surface electromyography (sEMG) based silent speech recognition interface have resulted in significant advances in the hardware, software and algorithmic components of the system. In this paper, we report our algorithmic progress, specifically: sEMG feature extraction, parameter optimization, advances in sEMG acoustic modeling, and sEMG sensor set reduction. The key findings are: 1) the gold-standard parameters for acoustic speech feature extraction are far from optimum for sEMG parameterization, 2) advances in state-of-the-art speech modeling can be leveraged to significantly enhance the continuous sEMG silent speech recognition accuracy, and 3) the number of sEMG sensors can be reduced by half with little impact on the final recognition accuracy, and the optimum sensor subset can be selected efficiently based on basic monophone HMM modeling.

Opti-Speech: A Real-Time, 3D Visual Feedback System for Speech Training

William Katz, Thomas F. Campbell, Jun Wang, Eric Farrar, J. Coleman Eubanks, Arvind Balasubramanian, Balakrishnan Prabhakaran, Rob Rennaker; University of Texas at Dallas, USA

We describe an interactive 3D system to provide talkers with real-time information concerning their tongue and jaw movements.
during speech. Speech movement is tracked by a magnetometer system (Wave; NDI, Waterloo, Ontario, Canada). A customized interface allows users to view their current tongue position (represented as an avatar consisting of flesh-point markers and a modeled surface) placed in a synchronously moving, transparent head. Subjects receive augmented visual feedback when tongue sensors achieve the correct place of articulation. Preliminary data obtained for a group of adult talkers suggest this system can be used to reliably provide real-time feedback for American English consonant place of articulation targets. Future studies, including tests with communication disordered subjects, are described.

Across-Speaker Articulatory Normalization for Speaker-Independent Silent Speech Recognition
Jun Wang¹, Ashok Samal², Jordan R. Green³; ¹University of Texas at Dallas, USA; ²University of Nebraska-Lincoln, USA; ³MGH Institute of Health Professions, USA

Silent speech interfaces (SSIs), which recognize speech from articulatory information (i.e., without using audio information), have the potential to enable persons with laryngectomy or a neurological disease to produce synthesized speech with a natural sounding voice using their tongue and lips. Current approaches to SSIs have largely relied on speaker-dependent recognition models to minimize the negative effects of talker variation on recognition accuracy. Speaker-independent approaches are needed to reduce the large amount of training data required from each user; only limited articulatory samples are often available for persons with moderate to severe speech impairments, due to the logistic difficulty of data collection. This paper reported an across-speaker articulatory normalization approach based on Procrustes matching, a bidimensional regression technique for removing translational, scaling, and rotational effects of spatial data. A dataset of short functional sentences was collected from seven English talkers. A support vector machine was then trained to classify sentences based on normalized tongue and lip movements. Speaker-independent classification accuracy (tested using leave-one-subject-out cross validation) improved significantly, from 68.63% to 95.90%, following normalization. These results support the feasibility of a speaker-independent SSI using Procrustes matching as the basis for articulatory normalization across speakers.

Conversion from Facial Myoelectric Signals to Speech: A Unit Selection Approach
Marlene Zahner, Matthias Janke, Michael Wand, Tanja Schultz; KIT, Germany

This paper reports on our recent research on surface electromyographic (EMG) speech synthesis: a direct conversion of the EMG signals of the articulatory muscle movements to the acoustic speech signal. In this work we introduce a unit selection approach which compares segments of the input EMG signal to a database of simultaneously recorded EMG/audio unit pairs and selects the best matching audio unit based on target and concatenation cost, which will be concatenated to synthesize an acoustic speech output. We show that this approach is feasible to generate a proper speech output from the input EMG signal. We evaluate different properties of the units and investigate what amount of data is necessary for an initial transformation. Prior work on EMG-to-speech conversion used a frame-based approach from the voice conversion domain, which struggles with the generation of a natural F0 contour. This problem may also be tackled by our unit selection approach.

Towards Real-Life Application of EMG-Based Speech Recognition by Using Unsupervised Adaptation
Michael Wand, Tanja Schultz; KIT, Germany

This paper deals with a Silent Speech Interface based on Surface Electromyography (EMG), where electrodes capture the electric activity generated by the articulatory muscles from a user’s face in order to decode the underlying speech, allowing speech to be recognized even when no sound is heard or created. So far, most EMG-based speech recognizers described in literature do not allow electrode reattachment between system training and usage, which we consider unsuitable for practical applications. In this study we report on our research on unsupervised session adaptation: A system is pre-trained with data from multiple recording sessions and then adapted towards the current recording session using data accruable during normal use, without requiring a time-consuming specific enrollment phase. We show that considerable accuracy improvements can be achieved with this method, paving the way towards real-life applications of the technology.

Simple Gesture-Based Error Correction Interface for Smartphone Speech Recognition
Yuan Liang¹, Koji Iwano², Koichi Shinoda¹; ¹Tokyo Institute of Technology, Japan; ²Tokyo City University, Japan

Conventional error correction interfaces for speech recognition require a user to first mark an error region and choose the correct word from a candidate list. Taking the user’s effort and the limited user interface available in a smartphone into account, this operation should be simpler. In this paper, we propose an interface where users mark the error region once, and then the word will be replaced by another candidate. Assuming that the words preceding/succeeding the error region are validated by the user, we search the Web n-grams for long word sequences matched to such a context. The acoustic features of the error region are also utilized to rerank the candidate words. The experimental result proved the effectiveness of our method. 30.2% of the error words were corrected by a single operation.

Keynote 3: Lori Lamel
Garnet 213-218
13:30 - 14:30, Tuesday 16 September 2014
Chair: Isabel Trancoso

Language Diversity: Speech Processing in a Multi-Lingual Context
Lori Lamel; LIMSI, France

Speech processing encompasses a variety of technologies that automatically process speech for some downstream processing. These technologies include identifying the language or dialect spoken, the person speaking, what is said and how it is said. The downstream processing may be limited to a transcription or to a transcription enhanced with additional metadata, or may be used to carry out an action or interpreted within a spoken dialog system or more generally for analytics. With the availability of large spoken multimedia or multimodal data there is growing interest in using such technologies to provide structure and random access to particular segments. Automatic tools can also serve to annotate large corpora for exploitation in linguistic studies of spoken

NOTES
language, such as acoustic-phonetics, pronunciation variation and diachronic evolution, permitting the validation of hypotheses and models. In this talk I will present some of my experience with speech processing in multiple languages, drawing upon progress in the context of several research projects, most recently the Quaero program and the IARPA Babel program, both of which address the development of technologies in a variety of languages, with the aim to some highlight recent research directions and challenges.

Oral Session 17: Normalization and Discriminative Training Methods
Garnet 213-218
15:00 – 17:00, Tuesday 16 September 2014
Chairs: Jean-Luc Gauvain and Geoffrey Zweig

Normalization of ASR Confidence Classifier Scores via Confidence Mapping
Kshitiz Kumar, Chaojun Liu, Yifan Gong; Microsoft, USA
Tue-O-17-1

Speech recognition confidence classifier (CC) score quantitatively represents the correctness of decoded utterances in a (0,1] range. We associate an operating threshold with the classifier and accept recognitions with scores greater than the threshold. Speech developers may set their own threshold but often an acoustic model (AM) or CC update alters the correct-accept (CA) vs. false-accept (FA) profile, necessitating a threshold reselection. This is specifically a problem when, (a) threshold is hardcoded with a shipped hardware or software, (b) developers may not have expertise for threshold tuning, (c) tuning isn’t cost-effective and may need to be done often.

To our knowledge, our work is the first to present this practical and interesting problem of avoiding threshold reselection and proposes novel confidence-mapping-based techniques to improve or retain both CA and FA at previously set thresholds. We propose and evaluate, (a) histogram-based mapping, (b) polynomial-fitting, (c) tanh-fitting, based methods to map confidences associated with false-recognitions and discuss their issues and benefits. In our tests, all of the above mapping methods fix the mean regression in CA of 21% to a gain to 1–2%, with tanh-mapping providing the best CA and FA tradeoff in our tests.

Neural Network Phone Duration Model for Speech Recognition
Tanel Alumäe; Tallinn University of Technology, Estonia
Tue-O-17-2

In this paper, we describe a novel phone duration model that is used to improve the accuracy of a large vocabulary speech recognition system based on state-of-the-art speaker-adapted DNN acoustic models. The duration model calculates the probability density function of phone duration from phone’s contextual features using a neural network which is then applied for word lattice rescoring. Experimental results are given for Estonian, English and Finnish transcription tasks. An absolute word error rate reduction of 0.8–1.4% is observed across all evaluation sets.

Sequence Discriminative Distributed Training of Long Short-Term Memory Recurrent Neural Networks
Hasim Sak, Oriol Vinyals, Georg Heigold, Andrew Senior, Erik McDermott, Rajat Monga, Mark Mao; Google, USA
Tue-O-17-3

We recently showed that Long Short-Term Memory (LSTM) recurrent neural networks (RNNs) outperform state-of-the-art deep neural networks (DNNs) for large scale acoustic modeling where the models were trained with the cross-entropy (CE) criterion. It has also been shown that sequence discriminative training of DNNs initially trained with the CE criterion gives significant improvements. In this paper, we investigate sequence discriminative training of LSTM RNNs in a large scale acoustic modeling task. We train the models in a distributed manner using asynchronous stochastic gradient descent optimization technique. We compare two sequence discriminative criteria — maximum mutual information and state-level minimum Bayes risk, and we investigate a number of variations of the basic training strategy to better understand issues raised by both the sequential model, and the objective function. We obtain significant gains over the CE trained LSTM RNN model using sequence discriminative training techniques.

Beyond Cross-Entropy: Towards Better Frame-Level Objective Functions for Deep Neural Network Training in Automatic Speech Recognition
Zhen Huang¹, Jinyu Li², Chao Weng¹, Chin-Hui Lee¹; ¹Georgia Institute of Technology, USA; ²Microsoft, USA
Tue-O-17-4

We propose two approaches for improving the objective function for the deep neural network (DNN) frame-level training in large vocabulary continuous speech recognition (LVCSR). The DNNs used in LVCSR are often constructed with an output layer with softmax activation and the cross-entropy objective function is always employed in the frame-leveling training of DNNs. The pairing of softmax activation and cross-entropy objective function contributes much to the success of DNN. The first approach developed in this paper improves the cross-entropy objective function by boosting the importance of the frames for which the DNN model has low target predictions (low target posterior probabilities) and the second one considers jointly minimizing the cross-entropy and maximizing the log posterior ratio between the target senone (tied-triphone states) and the most competing one. Experiments on Switchboard task demonstrate that the two proposed methods can provide 3.1% and 1.5% relative word error rate (WER) reduction, respectively, against the already very strong conventional cross-entropy trained DNN system.

A Comparison of Training Approaches for Discriminative Segmental Models
Hao Tang, Kevin Gimpel, Karen Livescu; Toyota Technological Institute at Chicago, USA
Tue-O-17-5

Segmental models such as segmental conditional random fields have had some recent success in lattice rescoring for speech recognition. They provide a flexible framework for incorporating a wide range of features across different levels of units, such as phones and words. However, such models have mainly been trained by maximizing conditional likelihood, which may not be the best proxy for the task loss of speech recognition. In addition, there has been little work on designing cost functions as surrogates for the word error rate. In this paper, we investigate various losses
and introduce a new cost function for training segmental models. We compare lattice rescoring results for multiple tasks and also study the impact of several choices required when optimizing these losses.

**Asynchronous Stochastic Optimization for Sequence Training of Deep Neural Networks: Towards Big Data**

**Erik McDermott, Georg Heigold, Pedro J. Moreno, Andrew Senior, Michiel Bacchiani; Google, USA**

Tue-O-17.6

Previous work presented a proof of concept for sequence training of deep neural networks (DNNs) using asynchronous stochastic optimization, mainly focusing on a small-scale task. The approach offers the potential to leverage both the efficiency of stochastic gradient descent and the scalability of parallel computation. This study presents results for four different voice search tasks to confirm the effectiveness and efficiency of the proposed framework across different conditions: amount of data (from 60 hours to 20,000 hours), type of speech (read speech vs. spontaneous speech), quality of data (supervised vs. unsupervised data), and language. Significant gains over baselines (DNNs trained at the frame level) are found to hold across these conditions. The experimental results are analyzed, and additional practical details for the approach are provided. Furthermore, different sequence training criteria are compared.

**Oral Session 18: Paralinguistic and Extralinguistic Information**

Peridot 202-203

15:00 - 17:00, Tuesday 16 September 2014

Chairs: Nick Campbell and Julien Epps

**Detection of Children’s Paralinguistic Events in Interaction with Caregivers**

Hrishikesh Rao1, Jonathan C. Kim1, Mark A. Clements1, Agata Rozga1, Daniel S. Messinger2;
1Georgia Institute of Technology, USA; 2University of Miami, USA

Tue-O-18-1

Paralinguistic cues in children’s speech convey the child’s affective state and can serve as important markers for the early detection of autism spectrum disorder (ASD). In this paper, we detect paralinguistic events, such as laughter and fussing/crying, along with toddlers’ speech from the Multi-modal Dyadic Behavior Dataset (MMDB). We use both spectral and prosodic acoustic features selected using a combination of filter and wrapper-based methods. The classification accuracy using a support vector machine with a linear kernel for detecting laughter in children’s speech was 77.87% and that for fussing/crying was 79.37%. A temporary classification scheme for detecting laughter, fussing/crying, and speech yielded an accuracy of 69.73%. To test for the generalization of the approach for detecting fussing/crying, we used recordings from the Strange Situation protocol, which is used to observe attachment behavior between an infant and a parent. Using a cross-corpus testing set for detecting fussing/crying, we obtained a detection accuracy of 71.6%. These results indicate that the selected acoustic features are capable of discriminating children’s laughter, fussing/crying, and speech and the algorithms generalize well to a dataset consisting of paralinguistic cues of a different age group, infants (12-18 months of age), gathered in a different context.

**Age and Rhythmic Variations: A Study on Italian**

**Massimo Pettorino, Elisa Pellegrino; Universität di Napoli “L’Orientale”, Italy**

Tue-O-18-2

Age-related variations in the speech production mechanism affecting acoustic cues such as pitch, speech rate, formant frequencies have been extensively investigated. Changes in speech rhythm and in utterance composition in terms of vowel and consonant proportions are, instead, scarcely examined. Given the relevance of the vowel percentage (%V) in the utterance and of the interval between two vowel onset points (VtoV) for the perception of language rhythm, this study aims to investigate the relationship between rhythmic variations and speaker age. It also attempts to determine whether %V can be affected by different speech rates. Four young adult and four old speakers of Italian read out loud at the same level of energy at four different speech rates. The whole corpus was segmented in vocalic and consonantal portions and in VtoV intervals. The analysis results have shown that aged voice accompanies with significant increase in %V, despite speech rate variations. With advancing age, Italian speech tends to shift from the isosyllabic towards the isomoraic rhythmic pattern.

**Probabilistic Acoustic Volume Analysis for Speech Affected by Depression**

Nicholas Cummins1, Vidhyasaharan Sethu1, Julien Epps1, Jarek Krajewski2, 1University of New South Wales, Australia; 2Universität Wuppertal, Germany

Tue-O-18-4

Alterations in speech motor control in depressed individuals have been found to manifest as a reduction in spectral variability. In this paper we present a novel method for measuring acoustic volume — a model-based measure that is reflective of this decrease in spectral variability — and assess the ability of features resulting from this measure for indexing a speaker’s level of depression. A Monte Carlo approximation that enables the computation of this measure is also outlined in this paper. Results obtained using the AVEC 2013 Challenge Dataset indicate there is a statistically significant reduction in acoustic variation with increasing levels of speaker depression, and using features designed to capture this change it is possible to outperform a range of conventional spectral measures when predicting a speaker’s level of depression.

**Exploring Modulation Spectrum Features for Speech-Based Depression Level Classification**

Elif Bozkurt1, Orith Toledo-Ronen2, Alexander Sorin2, Ron Hoory2; 1 Koç Üniversitesi, Turkey; 2IBM Research Haifa, Israel

Tue-O-18-5

In this paper, we propose a Modulation Spectrum-based manageable feature set for detection of depressed speech. Modulation Spectrum (MS) is obtained from the conventional speech spectrogram by spectral analysis along the temporal trajectories of the acoustic frequency bins. While MS representation of speech provides rich and high-dimensional joint frequency information, extraction of discriminative features from it remains as an open question. We propose a lower dimensional representation, which first employs a Mel-frequency filterbank in the acoustic frequency domain and Discrete Cosine Transform in the modulation frequency domain, and then applies feature selection in both domains. We compare and fuse the proposed feature set with other complementary prosodic and spectral features at the feature and decision levels. In our experiments, we use Support Vector Machines for discriminating the depressed speech in a

**NOTES**
Automatic Modelling of Depressed Speech: Relevant Features and Relevance of Gender
Florian Höning 1, Anton Bathliner 1, Elmar Nöth 1, Sebastian Schneider 2, Jarek Krajewski 2; 1FAU Erlangen-Nürnberg, Germany; 2Universität Wuppertal, Germany

Depression is an affective disorder characterised by psychomotor retardation; in speech, this shows up in reduction of pitch (variation, range), loudness, and tempo, and in voice qualities different from those of typical modal speech. A similar reduction can be observed in sleepy speech (relaxation). In this paper, we employ a small group of acoustic features modelling prosody and spectrum that have been proven successful in the modelling of sleepy speech, enriched with voice quality features, for the modelling of depressed speech within a regression approach. This knowledge-based approach is complemented by and compared with brute-forcing and automatic feature selection. We further discuss gender differences and the contributions of (groups of) features both for the modelling of depression and across depression and sleepiness.

Excitation Source Features for Discrimination of Anger and Happy Emotions
P. Gangamohan, Sudarsana Reddy Kadiri, Suryakanth V. Gangashetty, B. Yegnanarayana; IIIT Hyderabad, India

Studies on the emotion recognition task indicate that there is confusion in discrimination among higher activation states like ‘anger’ and ‘happy’. In this study, features related to excitation source of speech are examined for discriminating ‘anger’ and ‘happy’ emotions. The objective is to explore the features which are independent of lexical content, language, channel and speaker. The features like strength of excitation from zero frequency filter and spectrum that have been proven successful in the modelling of depression and across depression and sleepiness.

Oral Session 19: Text Processing for Speech Synthesis
Peridot 204-205
15:00 – 17:00, Tuesday 16 September 2014
Chairs: Alan Black and Jerome Bellegarda

Encoding Linear Models as Weighted Finite-State Transducers
Ke Wu 1, Cyril Allauzen 2, Keith Hall 2, Michael Riley 2, Brian Roark 2; 1University of Maryland, USA; 2Google, USA

We present algorithms, implemented as an extension to the Openfst library, that yield a class of transducers that encode linear models for structured inference tasks like segmentation and tagging. This allows the use of general finite-state operations with such models. For instance, finite-state composition can be used to apply the model to lattice input (or other more general automata) and then the result automaton can be passed to subsequent processing such as general shortest path algorithms. We demonstrate the use of the library extension on grapheme-to-phoneme conversion, encoding multiple varieties of linear models for that task, and achieve solid PER/WER gains over previous best reported results on g2p conversion of a publicly available dataset (CMU).

Structured Soft Margin Confidence Weighted Learning for Grapheme-to-Phoneme Conversion
Keigo Kubo, Sakriani Sakti, Graham Neubig, Tomoki Toda, Satoshi Nakamura; NAIIST, Japan

In recent years, structured online discriminative learning methods using second order statistics have been shown to outperform conventional generative and discriminative models in the grapheme-to-phoneme (g2p) conversion task. However, these methods update the parameters by sequentially using N-best hypotheses predicted with the current parameters. Thus, the parameters appearing in early hypotheses are overfitted compared with those in later hypotheses. In this paper, we propose a novel method called structured soft margin confidence weighted learning, which extends multi-class confidence weighted learning to structured learning. The proposed method extends multi-class CW in two ways, allowing for improved robustness to overfitting: (1) regularization inspired by soft margin support vector machines, allowing for margin error, and (2) update using N-best hypotheses simultaneously and interdependently. In an evaluation experiment on the g2p conversion task, the proposed method improved over all other approaches in terms of phoneme error rate by a significant difference.

Unsupervised Language Filtering Using the Latent Dirichlet Allocation
Wei Zhang 1, Robert A.J. Clark 2, Yongyuan Wang 1; 1Ocean University of China, China; 2University of Edinburgh, UK

To automatically build from scratch the language processing component for a speech synthesis system in a new language a purified text corpora is needed where any words and phrases from other languages are clearly identified or excluded. When using found data and where there is no inherent linguistic knowledge of the language/languages contained in the data, identifying the pure data is a difficult problem.

We propose an unsupervised language identification approach based on Latent Dirichlet Allocation where we take the raw n-gram count as features without any smoothing, pruning or interpolation. The Latent Dirichlet Allocation topic model is reformulated for the language identification task and Collapsed Gibbs Sampling is used to train an unsupervised language identification model. We show that such a model is highly capable of identifying the primary language in a corpus and filtering out other languages present.

Generating Multiple-Accent Pronunciations for TTS Using Joint Sequence Model Interpolation
BalaKrishna Kolluru, Vincent Wan, Javier Latorre, Kayoko Yanagisawa, Mark J.F. Gales; Toshiba Research Europe, UK

Standard grapheme-to-phoneme (G2P) systems are trained using a
homogeneous lexicon, for example one associated with a particular accent. In practice, a synthesis system may be required to handle multiple accents. Furthermore, a speaker rarely has a pure accent; accents vary continuously within and between regions of a country. Generating phonetic sequences for each accent is possible, but combining them to yield a single synthesis pronunciation is highly challenging. To address this problem, this paper considers a space of accents. The bases for these spaces are defined by statistical G2P models in the form of graphone models. A linear combination of these models define the accent space. By selecting a point in this continuous space, it is possible to specify the accent for an individual speaker. The performance of this approach is evaluated using an accent space defined by American, Scottish and British English. By moving around the accent space, it is shown that it is possible to synthesize speech from all these accents as well as a range of intermediate points.

Using a Hybrid Approach to Build a Pronunciation Dictionary for Brazilian Portuguese

Gustavo Mendonça, Sandra Aluísio; Universidade de São Paulo, Brazil

This paper describes the method employed to build a machine-readable pronunciation dictionary for Brazilian Portuguese. The dictionary makes use of a hybrid approach for converting graphemes into phonemes, based on both manual transcription rules and machine learning algorithms. It makes use of a word list compiled from the Portuguese Wikipedia dump. Wikipedia articles were transformed into plain text, tokenized and word types were extracted. A language identification tool was developed to detect loanwords among data. Words’ syllable boundaries and stress were identified. The transcription task was carried out in a two-step process: i) words are submitted to a set of transcription rules, in which predictable graphemes (mostly consonants) are transcribed; ii) a machine learning classifier is used to predict the transcription of the remaining graphemes (mostly vowels). The method was evaluated through 5-fold cross-validation; results show a F1-score of 0.98. The dictionary and all the resources used to build it were made publicly available.

A Flexible Front-End for HTS

Matthew P. Aylett1, Rasmus Dall2, Arnab Ghoshal2, Gustav Eje Henter2, Thomas Merritt2; 1CereProc, UK; 2University of Edinburgh, UK

Parametric speech synthesis techniques depend on full context acoustic models generated by language front-ends, which analyse linguistic and phonetic structure. HTS, the leading parametric synthesis system, can use a number of different front-ends to generate full context models for synthesis and training. In this paper we explore the use of a new text processing front-end that has been added to the speech recognition toolkit Kaldi as part of an ongoing project to produce a new parametric speech synthesis system, Idlak. The use of XML specification files, a modular design, and modern coding and testing approaches, make the Idlak front-end ideal for adding, altering and experimenting with the contexts used in full context acoustic models. The Idlak front-end was evaluated against the standard Festival front-end in the HTS system. Results from the Idlak front-end compare well with the more mature Festival front-end (Idlak - 2.83 MOS vs Festival - 2.55 MOS), although a slight reduction in naturalness perceived by non-native English speakers can be attributed to Festival’s insertion of non-punctuated pauses.

Notes
Acoustic Properties of Shared Vowels in Bilingual Mandarin-English Children
Jing Yang, Robert Allen Fox; Ohio State University, USA

This study investigates L1-L2 interactions in relatively young Mandarin (L1)-English (L2) bilingual children through comparing their static and dynamic vowel acoustic features with those of age-matched monolingual children. Two groups of sequential bilingual children aged 5-6 years (one with low proficiency in English and the other with high proficiency in English) were recorded producing a set of words containing the shared vowels /a/, /i/, /u/ in both languages. Age-matched monolingual children only produced the words in their native language. F1 and F2 values were measured at 5 equidistant time locations. It is found that both groups of bilingual children showed distinctive vowel dispersion patterns and dynamic spectral change patterns from those of monolingual children. Low proficiency bilingual children showed an assimilatory process of L1 on L2 and high proficiency bilingual children showed an assimilatory process of L2 on L1.

Generating Segmental Foreign Accent
Maria Luisa García Lecumberri1, Roberto Barra-Chicote2, Rubén Pérez Ramón1, Junichí Yamagishi3, Martín Cooke4; 1Universidad del País Vasco, Spain; 2Universidad Politécnica de Madrid, Spain; 3University of Edinburgh, UK; 4Ikerbasque, Spain

For most of us, speaking in a non-native language involves deviating to some extent from native pronunciation norms. However, the detailed basis for foreign accent (FA) remains elusive, in part due to methodological challenges in isolating segmental from suprasegmental factors. The current study examines the role of segmental features in conveying FA through the use of a generative approach in which accent is localised to single consonantal segments. Three techniques are evaluated: the first requires a highly-proficiency bilingual to produce words with isolated accented segments; the second uses cross-splicing of context-dependent consonants from the non-native language into native words; the third employs hidden Markov model synthesis to blend voice models for both languages. Using English and Spanish as the native/non-native languages respectively, listener cohorts from both languages identified words and rated their degree of FA. All techniques were capable of generating accented words, but to differing degrees. Naturally-produced speech led to the strongest FA ratings and synthetic speech the weakest, which we interpret as the outcome of over-smoothing. Nevertheless, the flexibility offered by synthesising localised accent encourages further development of the method.

Differences of Pitch Profiles in Germanic and Slavic Languages
Bistra Andreeva1, Grażyna Demenko2, Bernd Möbius1, Frank Zimmerer1, Jeanin Jügler1, Magdalena Oleszkowicz-Popiel2; 1Universität des Saarlandes, Germany; 2Adam Mickiewicz University in Poznań, Poland

This study investigates cross-language differences in pitch range and variation in four languages from two language groups: English and German (Germanic) and Bulgarian and Polish (Slavic). The analysis is based on large multi-speaker corpora (48 speakers for Polish, 60 for each of the other three languages). Linear mixed models were computed that include various distributional measures of pitch level, span and variation, revealing characteristic differences across languages and between language groups. A classification experiment based on the relevant parameter measures (span, kurtosis and skewness values for pitch distributions for each speaker) succeeded in separating the language groups.

The Obligatory Contour Principle in African and European Varieties of French
Mathieu Avanzi1, Guri Bordal2, Gélace Nimbona3; 1LLF (UMR 7110), France; 2University of Oslo, Norway; 3Université catholique de Louvain, Belgium

This study provides evidence that the role of the Obligatory Contour Principle (OCP) in substrate languages is reflected in prosodic systems of contact varieties of French. We have compared two contact varieties: Central African French, a variety of French spoken by L1 speakers of Sango a lexical tone language where the constraint is not respected, and Burundi French, which is spoken by L1 speakers of Kirundi, a lexical tone language where the OCP plays an important role in the distribution of tones. Our data indicate that clashes are permitted in the former, but avoided in the latter.

Content Matching for Short Duration Speaker Recognition
Nicolas Scheffer, Yun Lei; SRI International, USA

This work attempts to tackle the problem of content mismatch for short duration speaker verification. Experiments are run on both text-dependent and text-independent protocols, where a larger amount of enrollment data is available in the latter. We recently proposed a framework based on a deep neural network that explicitly utilizes phonetic information, and showed increased performance on long duration utterances. We show how this new framework can also yield significant improvements for short duration. We then propose an innovative approach to perform content matching, i.e. transforming a text-independent trial into a text-dependent one by mining content from a speaker’s enrollment data to match the test utterance. We show how content matching can be effectively done at the statistics level to enable the use
of standard verification backends. Experiments — run on the RSR2015 and NIST SRE 2010 data sets — show relative improvements of 50% for cases where the content has been said during enrollment. With no significant improvements were observed for the general text-independent case, we believe that this work might pave the way for new research for speaker verification with very short utterances.

**Extended RSR2015 for Text-Dependent Speaker Verification Over VHF Channel**

Anthony Larcher, Kong Aik Lee, Pablo L. Sordo Martínez, Trung Hieu Nguyen, Bin Ma, Haizhou Li; A*STAR, Singapore

Text-dependent speaker verification over degraded radio channel is a challenging task. To better understand the research problem, the Institute for Infocomm Research (I2R) of Singapore has collected a corpus of voice recordings transmitted over marine VHF. Built as an extension of the RSR2015 database, the VHF-RSR2015 consists of recordings from 300 speakers of Part I of the RSR2015 database transmitted over VHF channel. Extending the RSR2015 database, we would like to facilitate the study of the VHF channel effect, therefore, keeping the acoustic environment the same as that of RSR2015. Performance benchmark of a text-dependent speaker verification engine is given as reference on the original RSR2015 database recorded at a sampling frequency of 16kHz, on a sub-sampled version of the same dataset at 8kHz and on the data after transmission through the VHF channel.

**Tandem Deep Features for Text-Dependent Speaker Verification**

Tianfan Fu, Yanmin Qian, Yuan Liu, Kai Yu; Shanghai Jiao Tong University, China

Although deep learning has been successfully used in acoustic modeling of speech recognition, it has not been thoroughly investigated and widely accepted for speaker verification. This paper describes an investigation of using various types of deep features in a Tandem fashion for text-dependent speaker verification. Three types of networks are used to extract deep features: restricted Boltzmann machine (RBM), phone discriminant and speaker discriminant deep neural network (DNN). Hidden layer outputs from these networks are concatenated with the original acoustic features and used in a GMM-UBM classifier. The systems with Tandem deep feature were evaluated on RSR2015, a short-term text dependent speaker verification task. Experiments showed that the best Tandem deep feature obtained more than 50% relative EER reduction over the traditional feature in a GMM-UBM framework.

**In-Domain versus Out-of-Domain Training for Text-Dependent JFA**

Patrick Kenny¹, Themos Stafylakis¹, M.J. Alam¹, Pierre Ouellet¹, Marcel Kockmann²; ¹CRIM, Canada; ²VoiceTrust, Canada

We propose a simple and effective strategy to cope with dataset shifts in text-dependent speaker recognition based on Joint Factor Analysis (JFA). We have previously shown how to compensate for lexical variation in text-dependent JFA by adapting the Universal Background Model (UBM) to individual passphrases. A similar type of adaptation can be used to port a JFA model trained on out-of-domain data to a given text-dependent task domain. On the RSR2015 test set we found that this type of adaptation gave essentially the same results as in-domain JFA training. To explore this idea more fully, we experimented with several types of JFA model on the CSLU speaker recognition dataset. Taking a suitably configured JFA model trained on NIST data and adapting it in the proposed way results in a 22% reduction in error rates compared with the GMM/UBM benchmark. Error rates are still much higher than those that can be achieved on the RSR2015 test set with the same strategy but cheating experiments suggest that if large amounts of in-domain training data are available, then JFA modelling is capable in principle of achieving very low error rates even on hard tasks such as CSLU.

**Domain Adaptation for Text Dependent Speaker Verification**

Hagai Aronowitz, Asaf Rendel; IBM Research Haifa, Israel

Recently we have investigated the use of state-of-the-art text-dependent speaker verification algorithms for user authentication and obtained satisfactory results mainly by using a fair amount of text-dependent development data from the target domain. In this work we investigate the ability to build high accuracy text-dependent systems using no data at all from the target domain. Instead of using target domain data, we use resources such as TIMIT, Switchboard, and NIST data. We introduce several techniques addressing both lexical mismatch and channel mismatch. These techniques include synthesizing a universal background model according to lexical content, automatic filtering of irrelevant phonetic content, exploiting information in residual supervectors (usually discarded in the i-vector framework), and inter-dataset variability modeling. These techniques reduce verification error significantly, and also improve accuracy when target domain data is available.

**Factor Analysis with Sampling Methods for Text Dependent Speaker Recognition**

Antonio Miguel¹, Jesús Villalba¹, Alfonso Ortega¹, Eduardo Lleida¹, Carlos Vaquero²; ¹Universidad de Zaragoza, Spain; ²AGNITIO, Spain

Factor analysis is a method for embedding high dimensional data into a lower dimensional factor space. When data are multimodal we use mixtures of factor analyzers (MFA), which assume statistically independent samples. In speaker recognition, samples are not independent because they depend on the speaker in the utterance. In joint factor analysis and i-vectors, the MFA latent factors are tied at different levels. For example, they can be tied for a segment to extract utterance level information. Tied MFA approaches usually present the drawback that computing the exact posterior of the hidden variables (component responsibilities and latent factors) is unfeasible. For JFA, the preferred approximation consists in computing the responsibilities given a speaker independent GMM and they are fixed during the rest of the process. That implies that the estimated responsibilities for a given sample are independent of the rest of the samples of the utterance not taking into account the shared speaker and channel. We present a novel approximation to jointly estimate responsibilities and latent factors based on sampling the latent factor space. This model differs from previous ones in the hidden variables and parameter estimation, and likelihood evaluation. This approach was tested on the RSR2015 database for text-dependent speaker recognition.
Pitch detection has important applications in areas of automatic speech recognition such as prosody detection, tonal language transcription, and general feature augmentation. In this paper we describe Pitcher, a new pitch tracking algorithm that correlates spectral information with a dictionary of waveforms each of which is designed to match signals with a given pitch value. We apply dynamic programming techniques on the resulting coefficient matrix to extract a smooth pitch contour while facilitating pitch halving and doubling transitions. We discuss the design of pitch atoms along with the various considerations for the pitch extraction process. We evaluate the performance of Pitcher on the PTDB database and compare its performance with three existing pitch tracking algorithms: YIN, IRAPT, and Swift. The performance of Pitcher consistently outperforms the other methods for low-pitched speakers and is comparable in performance to the best of the other three methods for high-pitched speakers.

Acoustic Features for Robust Classification of Mandarin Tones

Hongbing Hu, Stephen A. Zahorian, Peter Guzewich, Jiang Wu; Binghamton University, USA

For applications such as tone modeling and automatic tone recognition, smoothed F0 (pitch) all-voiced pitch tracks are desirable. Three pitch trackers that have been shown to give good accuracies for pitch tracking are YAAPT, YIN, and PRAAT. On tests with English and Japanese databases, for which ground truth pitch tracks are available by other means, we show that YAAPT has lower errors than YIN and PRAAT. We also experimentally compare the effectiveness of the three trackers for automatic classification of Mandarin tones. In addition to F0 tracks, a compact set of low-frequency spectral shape trajectories are used as additional features for automatic tone classification. A combination of pitch trajectories computed with YAAPT and spectral shape trajectories extracted from 800ms intervals for each tone results in tone classification accuracy of nearly 77%, a rate higher than human listeners achieve for isolated tonal syllables, and also higher than that obtained with the other two trackers.

Preservation of Lexical Tones in Singing in a Tone Language

Anastasia Karlsson, Håkan Lundström, Jan-Olof Svantesson; Lund University, Sweden

Lexical tones are important for expressing meaning and usually have high priority in tone languages. This can create conflicts with sentence intonation in spoken language and with melodic templates in singing since all of these are transmitted by pitch. The main question in this investigation is whether a language (in our case the Mon-Khmer language Kammu) with a simple two-tone system uses similar strategies for preserving lexical tones in singing and speech. We investigate the realization of lexical tones in a singing genre which can be described as recitation based on a partly predefined, though still flexible, melodic template. The contrast between High and Low tone is preserved, and is realized mainly at the beginning of the vowel. Apparently, the rest of the syllable rhyme serves either for strengthening the lexical contrast or for melodic purposes. Syllables are often reduplicated in singing, and the reduplicant ignores lexical tones. The preservation of lexical tones in Kammu singing, and their early timing close to the vowel onset, is very similar to what we have found for speech.
We present a novel method for determining how the perceptual organisation of simple alternating tone sequences is likely to occur in human listeners. By training a tensor model representation using features which incorporate both low-frequency modulation rate and phase, a set of components is learned. Test patterns are evaluated against each component to produce a score indicating how well the pattern fits the component. The combination of these scores provides a measure of how well the pattern fits the overall model. We find that for the basic streaming paradigm tested, our proposed model and method is able to correctly predict either segregation or integration in the majority of cases.

Detection of Vowel Onset Points in Voiced Aspirated Sounds of Indian Languages
BiswaJit Dev Sarma, S.R.M. Prasanna; IIT Guwahati, India
Tue-P-13-7, Poster

Vowel onset point (VOP) is defined as the instant at which onset of vowel takes place. Accurate detection of VOP is useful in many applications like syllable unit recognition, end-point detection, speaker verification etc. Manually and automatically locating VOPs accurately in case of voiced aspirated (VA) sounds is found to be difficult and ambiguous. This is due to the complex nature of the speech signal waveform around the VOP. This work addresses this issue and a manual marking approach using electroglottograph (EGG) signal is described which accurately marks the VOPs without any ambiguity. The knowledge derived from this manual analysis is transformed into an automatic method for the detection of VOPs in VA sounds. An automatic method is proposed using both source and vocal tract information. VOP detection accuracy of the proposed method is found to be significantly higher than some of the state of the art techniques.

Accuracy Evaluation of Esophageal Voice Analysis Based on Automatic Topology Generated-Voicing Source HMM
Akira Sasou; AIST, Japan
Tue-P-13-8, Poster

An Auto-Regressive eXogenous (ARX) model combined with descriptive models of the glottal source waveform has been adopted to more accurately separate the vocal tract and the voicing source. However, these methods cannot be easily applied to the analysis of voices uttered by different speech production methods, such as esophageal voice. We previously proposed the Voicing Source Hidden Markov Model (VS-HMM) and an accompanying parameter estimation method. The states of the VS-HMM were concatenated in a ring topology to represent the periodicity of the glottal source. We refer to the model combining the VS-HMM with an Auto-Regressive (AR) filter as AR-HMM. In this paper, we extend the conventional AR-HMM with its fixed ring topology to automatically generate the optimum topology for the VS-HMM using the Minimum Description Length-based Successive State Splitting (MDL-SSS) algorithm in order to simultaneously and accurately estimate the vocal tract and voicing source based on a voice excited by an unknown, aperiodic voicing source such as an esophageal voice.

Experiment results using synthesized pseudo-esophageal voices confirmed that the proposed AR-HMM approach can separate the vocal tract characteristics and the voicing source more accurately than the conventional AR-HMM with its fixed-ring topology or the LP method.

Audio Watermarking Based on Multiple Echoes Hiding for FM Radio
Xuejun Zhang, Xiang Xie; BIT, China
Tue-P-13-9, Poster

An audio watermarking system based on multiple echoes hiding is designed for frequency modulation (FM) broadcasting. The embedded watermarking information of 16 bits can be recovered from the receiver by recording the broadcast for any 1 second. On the premise of guaranteeing the high imperceptibility and robustness, the multiple echoes scheme we proposed has a higher capacity. The proposed system is tested on a semi-physical platform with the real channel transmission and the recovery rate reaches higher than 98%. Some attacks, such as white noise, filtering, re-sampling, adding an echo and re-quantization, are used to test the robustness. The subjective test result shows that the system has a good imperceptibility.

Poster Session 14: Cross-Lingual and Adaptive Language Modeling
Max Atria Gallery
15:00 - 17:00, Tuesday 16 September 2014
Chair: Tara Sainath

Development of Bilingual ASR System for MediaParl Corpus
Petr Motlicek1, David Imseng1, Milos Cernak1, Namhoon Kim2; 1Idiap Research Institute, Switzerland; 2Samsung Electronics, Korea
Tue-P-14-1, Poster

The development of an Automatic Speech Recognition (ASR) system for the bilingual MediaParl corpus is challenging for several reasons: (1) reverberant recordings, (2) accented speech, and (3) no prior information about the language. In that context, we employ frequency domain linear prediction-based (FDLP) features to reduce the effect of reverberation, exploit bilingual deep neural networks applied in Tandem and hybrid acoustic modeling approaches to significantly improve ASR for accented speech and develop a fully bilingual ASR system using entropy-based decoding-graph selection. Our experiments indicate that the proposed bilingual ASR system performs similar to a language-specific ASR system if approximately five seconds of speech are available.

Investigation of Cross-Lingual Bottleneck Features in Hybrid ASR Systems
Jie Li, Rong Zheng, Bo Xu; Chinese Academy of Sciences, China
Tue-P-14-2, Poster

In this work, we investigate cross-lingual BN features in hybrid ASR systems under a two-level DNNs framework. The first-level DNNs are bottleneck feature extractors and the second-level DNNs serve as not only acoustic models but also feature combination modules. Different feature configurations, including the bottleneck dimensionality, the need of delta processing and the necessity of concatenation with standard features of target-language, are first
studied. Further experiments are done to evaluate the cross-lingual generalization in a more holistic manner using optimized features. We then analyze the effects of adding more training data on the BN feature extractors. Performance improvement can be obtained when more data available. Finally, two different approaches of utilizing data from non-target languages are experimentally compared. It is shown that these two approaches have similar performance with each other, and the two-level DNNs architecture benefits from either of them.

Language Identification of Individual Words with Joint Sequence Models

Oluwapelumi Giwa, Marelie H. Davel; North-West University, South Africa

Within a multilingual automatic speech recognition (ASR) system, knowledge of the language of origin of unknown words can improve pronunciation modelling accuracy. This is of particular importance for ASR systems required to deal with code-switched speech or proper names of foreign origin. For words that occur in the language model, but do not occur in the pronunciation lexicon, text-based language identification (T-LID) of a single word in isolation may be required. This is a challenging task, especially for short words. We motivate for the importance of accurate T-LID in speech processing systems and introduce a novel way of applying Joint Sequence Models to the T-LID task. We obtain competitive results on a real-world 4-language task: for our best JSM system, an F-measure of 97.2% is obtained, compared to a F-measure of 95.2% obtained with a state-of-the-art Support Vector Machine (SVM).

Audio-to-Text Alignment for Speech Recognition with Very Limited Resources

Xavier Anguera, Jordi Luque, Ciro Gracia; Telefónica I+D, Spain

In this paper we present our latest investigations of language modeling for Code-Switching. Since there is only little text material for Code-Switching speech available, we integrate syntactic and semantic features into the language modeling process. In particular, we use part-of-speech tags, language identifiers, Brown word clusters and clusters of open class words. We develop factored language models and convert recurrent neural network language models into backoff language models for an efficient usage during decoding. A detailed error analysis reveals the strengths and weaknesses of the different language models. When we interpolate the models linearly, we reduce the perplexity by 15.6% relative on the SEAME evaluation set. This is even slightly better than the result of the unconverted recurrent neural network. We also combine the language models during decoding and obtain a mixed error rate reduction of 4.4% relative on the SEAME evaluation set.

Data Augmentation, Feature Combination, and Multilingual Neural Networks to Improve ASR and KWS Performance for Low-Resource Languages

Zoltán Tüské, Pavel Golik, David Nolden, Ralf Schlüter, Hermann Ney; RWTH Aachen University, Germany

This paper presents the progress of acoustic models for low-resource languages (Assamese, Bengali, Haitian Creole, Lao, Zulu) developed within the second evaluation campaign of the IARPA Babel project. This year, the main focus of the project is put on training high-performing automatic speech recognition (ASR) and keyword search (KWS) systems from language resources limited to about 10 hours of transcribed speech data. Optimizing the structure of Multilayer Perceptron (MLP) based feature extraction and switching from the sigmoid activation function to rectified

NOTES
This paper presents a flexible topic model based on the nested Indian buffet process (nIBP). The flexibility is achieved by relaxing three constraints: (1) number of topics is fixed, (2) topics are independent, and (3) topic hierarchy for a document is limited by a single tree path. Bayesian nonparametric learning is conducted to build a tree model where the number of topics and the topic hierarchies are automatically learnt from the given data. In particular, we propose the nIBP to construct the topic mixture model for representation of heterogeneous documents where the mixture components are flexibly selected from tree nodes or dishes that a document or customer chooses in Indian buffet process. The selection is performed in a nested and hierarchical manner. The experiments on document representation show the benefits of using the proposed nIBP.

Automated Closed Captioning for Russian Live Broadcasting

K. Levin¹, I. Ponomareva¹, A. Bulusheva¹, G. Chernykh², I. Medennikov¹, N. Merkin¹, A. Prudnikov¹, Natalia Tomashenko¹,¹ Speech Technology Center, Russia;²Saint-Petersburg State University, Russia

The paper describes a hardware-software system for real-time closed captioning of Russian live TV broadcasts. The use of respeaking technology enabled us to create an ASR system with WER not exceeding 5.5%. Editing closed captions in real time further reduces WER down to 0.2%. In the paper we report some advances in LMs for a highly inflected language and also in using morphological rescoring of the decoder word lattice. We propose a solution of the punctuation problem and effective methods of real-time editing of ASR results. This system was successfully used during paralympic games in Sochi for live web-broadcasting on russiasport.ru. We are reporting work in progress and are planning to achieve even better ASR accuracy in the course of the next year.
by including English words. Hence, the Mandarin ASR system becomes capable to recognize English words without retraining or re-estimation of the acoustic model parameters. Using the lexicon that derived from the proposed rules, the ASR performance of Mandarin English mixed speech is improved without harming the accuracy of Mandarin only speech. The proposed lexical rules are generalized and they can be directly applied to unseen English words.

**Pronunciation Learning for Named-Entities Through Crowd-Sourcing**

Attapol T. Rutherford¹, Fuchun Peng², Françoise Beaujays²; ¹Brandeis University, USA; ²Google, USA

Obtaining good pronunciations for named-entities poses a challenge for automated speech recognition because named-entities are diverse in nature and origin, and new entities come up every day. In this paper, we investigate the feasibility of learning named-entity pronunciations using crowd-sourcing. By collecting audio samples from non-linguistic-expert speakers with Mechanical Turk and learning from them, we can quickly derive pronunciations that are more accurate in speech recognition tests than manual pronunciations generated by linguistic experts. Compared to traditional approaches of generating pronunciations, this new approach proves to be cheap, fast, and quite accurate.

**Pronunciation Variation in Read and Conversational Austrian German**

Barbara Schuppler¹, Martine Adda-Decker², Juan A. Morales-Cordovilla¹; ¹Technische Universität Graz, Austria; ²LPP (UMR 7018), France

This paper presents the first large-scale analysis of pronunciation variation in conversational Austrian German. Whereas for the varieties of German spoken in Germany, conversational speech has been given noticeable attention in the fields of linguistics and automatic speech recognition, for conversational Austrian there is a lack in speech resources and tools as well as linguistic and phonetic studies. Based on the recently collected GRASS corpus, we provide (methods for the creation of) a pronunciation dictionary and (tools for the creation of) broad phonetic transcriptions for Austrian German. Subsequently, we present a comparative analysis of the occurrence of phonological and reduction rules in read and conversational speech. We find that whereas some rules are specific for the Austrian Standard variant and thus occur in both speech styles (e.g., the realization of /z/ as [s]), other rules are specific for conversational speech (e.g., the realization of /a/ as [o]). Overall, our results show that less words are produced with the citation form for conversational Austrian German (37.8%) than for other languages of the same style (e.g., Dutch conversations: 56%).

**Discriminative Pronunciation Modeling for Dialectal Speech Recognition**

Maider Lehr¹, Kyle Gorman¹, Izhak Shafran²; ¹Oregon Health & Science University, USA; ²Google, USA

Speech recognizers are typically trained with data from a standard dialect and do not generalize to non-standard dialects. Mismatch mainly occurs in the acoustic realization of words, which is represented by acoustic models and pronunciation lexicon. Standard techniques for addressing this mismatch are generative in nature and include acoustic model adaptation and expansion of lexicon with pronunciation variants, both of which have limited effectiveness. We present a discriminative pronunciation model whose parameters are learned jointly with parameters from the language models. We tease apart the gains from modeling the transitions of canonical phones, the transduction from surface to canonical phones, and the language model. We report experiments on African American Vernacular English (AAVE) using NPR’s StoryCorps corpus. Our models improve the performance over the baseline by about 2.1% on AAVE, of which 0.6% can be attributed to the pronunciation model. The model learns the most relevant phonetic transformations for AAVE speech.

**The Goodness of Pronunciation Algorithm Applied to Disordered Speech**

Thomas Pellegrini¹, Lionel Fontan¹, Julie Mauleclair¹, Jérôme Farinas¹, Marina Robert²; ¹IRIT, France; ²Université Paris Ouest, France

In this paper, we report on a study with the aim of automatically detecting phoneme-level mispronunciations in 32 French speakers suffering from unilateral facial palsy at four different clinical severity grades. We sought to determine if the Goodness of Pronunciation (GOP) algorithm, which is commonly used in Computer-Assisted Language Learning systems to detect learners’ individual errors, could also detect segmental deviations in disordered speech. For this purpose, speech read by the 32 speakers was aligned and GOP scores were computed for each phone realization. The highest scores, which indicate large dissimilarities with standard phone realizations, were obtained for the most severely impaired speakers. The corresponding speech subset was manually transcribed at phone-level. 8.3% of the phones differed from standard pronunciations extracted from our lexicon. The GOP technique allowed to detect 70.2% of mispronunciations with an equal rate of about 30% of false rejections and false acceptances. The phone substitutions detected by the algorithm confirmed that some of the speakers have difficulties to produce bilabial plosives, and showed that other sounds such as sibilants are prone to mispronunciation. Another interesting finding was the fact that speakers diagnosed with a same pathology grade do not necessarily share the same pronunciation issues.

**Using Deep Neural Networks to Improve Proficiency Assessment for Children English Language Learners**

Angeliki Metallinou, Jian Cheng; Pearson, USA

We investigated the use of context-dependent deep neural network hidden Markov models, or CD-DNN-HMMs, to improve speech recognition performance for a better assessment of children English language learners (ELLs). The ELL data used in the present study was obtained from a large language assessment project administered in schools in a U.S. state. Our DNN-based speech recognition system, built using rectified linear units (ReLU), greatly outperformed recognition accuracy of Gaussian mixture models (GMM)-HMMs, even when the latter models were trained with eight times more data. Large improvement was observed for cases of noisy and/or unclear responses, which are common in ELL children speech. We further explored the use of content and manner-of-speaking features, derived from the speech recognizer output, for estimating spoken English proficiency levels. Experimental results show that the DNN-based recognition approach achieved 31% relative WER reduction when compared to GMM-HMMs. This further improved the quality of the extracted features and final
Alignment of Spoken Utterances with Slide Content for Easier Learning with Recorded Lectures Using Structured Support Vector Machine (SVM)

Han Lu, Sheng-syan Shen, Sz-Rung Shiang, Hung-yi Lee, Lin-shan Lee; National Taiwan University, Taiwan

This paper reports the first known effort to automatically align the spoken utterances in recorded lectures with the content of the slides used. Such technologies will be very useful in Massive Open Online Courses (MOOCs) and various recorded lectures as well as many other applications. We propose a set of approaches considering the problem that words helpful for such alignment are sparse and noisy, and the assumption that the presentation of a slide is usually smooth and top-down across the slide. This includes utterance clustering, entropy-based word filtering, reliability-propagated word-based matching, and the structured support vector machine (SVM) learning from local and global features. Initial experimental results with the lectures in a course offered in National Taiwan University showed very encouraging results as compared to the baseline approaches.

A Preliminary Study on ASR-Based Detection of Chinese Mispronunciation by Japanese Learners

Richeng Duan, Jinsong Zhang, Wen Cao, Yanlu Xie; BLCU, China

Detecting mispronunciations produced by non-native speakers and providing detailed instructive feedbacks are desired in computer assisted pronunciation training system (CAPT), as it is helpful to L2 learners to improve their pronunciation more effectively. In this paper, we present our preliminary study on detecting phonetic segmental mispronunciations on account of the erroneous articulation tendencies, including the place of articulation and the manner of articulation. Through modeling and detecting these error patterns, feedbacks based on articulation-placement and articulation-manner could be given. Moreover, Japanese learners of Chinese are focused on in this study. The experimental results show that the approach can detect the mostly representative pronunciation errors moderately well, achieving a false rejection rate of 8.0% and a false acceptance rate 32.6%. The diagnostic accuracy is 86.0%.

Show and Tell Session 1

15:00 – 17:00, Tuesday 16 September 2014
Chair: Alvina Goh Siew Wee

3D Tongue Motion Visualization Based on Ultrasound Image Sequences

Kele Xu1, Yin Yang2, A. Jaumard-Hakoun1, Martine Adda-Decker1, A. Amelot3, S.K. Al Kork1, L. Crevier-Buchman3, P. Chawah2, G. Dreyfus4, T. Fux3, C. Pillot-Loiseau1, P. Roussel3, M. Stone3, B. Denby1; 1UPMC, France; 2University of New Mexico, USA; 3University of Maryland, USA

The article proposes a real-time technique for visualizing tongue motion driven by ultrasound image sequences. Local feature description is used to follow characteristic speckle patterns in a set of mid-sagittal contour points in an ultrasound image sequence, which are then used as markers for describing movements of the tongue. A 3D tongue model is subsequently driven by the motion data extracted from the ultrasound image sequences. The “modal warping” technique is used for real-time tongue deformation visualization. The resulting system will be useful in a variety of domains including speech production study, articulation training, educational scenarios, etc. Some parts of the interface are still being developed; we will show preliminary results in the demonstration.

Listen with Your Skin: Aerotak Speech Perception Enhancement System

Donald Derrick, Tom De Rybel, Greg A. O’Beirne, Jennifer Hay; University of Canterbury, New Zealand

Here we introduce Aerotak: A system for audio analysis and perception enhancement that allows speech perceivers to listen with their skin. The current system extracts unvoiced portions of an audio signal representative of turbulent air-flow in speech. It stores the audio signal in the left channel of a stereo audio output, and the air flow signal is stored in the right channel. The stored audio is used to drive a conversion unit that splits the left audio channel into a headphone out (to both ears) and right channel air pump drive signal to a piezoelectric pump that is mounted to the headphones. We have shown, using two-way forced-choice experiments, that the system enhances perception of voiceless stops and voiceless fricatives in noise such that 1 out of every 4 such words that would otherwise be missed will be heard correctly. We are currently conducting experiments on word identification while listening to a short-story, and are completing a stand-alone version of the Aerotak that works with real-time audio and from an embedded system. The short-story research and real-time system will be complete for InterSpeech 2014.

Speech Assistant System

László Czap; University of Miskolc, Hungary

A speech assistant system has been developed at the University of Miskolc in cooperation with the University of Debrecen granted by the European Union. The project aims to help training of deaf and hearing impaired people to speak. The idea of the project has come from a three-dimensional head model for articulation.
presentation, called “talking head” developed by the University of Miskolc, and an audio-visual transcoder for sound visualization developed by the University of Debrecen.

**Spoken Dialogue System for Restaurant Recommendation and Reservation**

Rafael E. Banchs, Seokhwan Kim; A*STAR, Singapore  
Show&Tell-1-7

This Show & Tell demo paper describes a multi-strategy spoken dialogue system for restaurant recommendation and reservation in Singapore. The system uses a three phased dialogue strategy for recommendation, selection and booking. Given its simple architecture, it can be easily adapted to deploy restaurant recommendation services for any city, as far as the required regional data is available.

**Interlingual Map Task Corpus Collection**

Hayakawa Akira, Nick Campbell, Saturnino Luz; Trinity College Dublin, Ireland  
Show&Tell-1-5

We present a prototype interlingual communication system that is being used to collect a corpus of task based dialogues between speakers of different languages. This corpus will be used to assess human reactions to an automated speech-to-speech translation system. In this demonstration we show how the HCRC Map Task can be adapted to support data collection in this interlingual environment, and how we used easily accessible speech and language technology for the rapid prototyping of the system used for data collection. An explanation of the nature and purpose of the data we are collecting is also presented.

**A Client Mobile Application for Chinese-Spanish Statistical Machine Translation**

Jordi Centelles, Marta R. Costa-jussà, Rafael E. Banchs; Universitat Politècnica de Catalunya, Spain; A*STAR, Singapore  
Show&Tell-1-6

This show and tell paper describes a client mobile application for Chinese-Spanish machine translation. The system combines a standard server-based statistical machine translation (SMT) system, which requires online operation, with different input modalities including text, optical character recognition (OCR) and automatic speech recognition (ASR). It also includes an index-based search engine for supporting off-line translation.

**LuciaWebGL: A New WebGL-Based Talking Head**

Alberto Benin, Piero Cosi, Giuseppe Riccardo Leone, Giulio Paci; CNR-ISTC, Italy  
Show&Tell-1-7

In this DEMO we present the first worldwide WebGL implementation of a talking head (LuciaWebGL), and also the first WebGL talking head running on iOS mobile devices (Apple iPhone and iPad).

**Crowdee: Mobile Crowdsourcing Micro-Task Platform for Celebrating the Diversity of Languages**

Babak Naderi, Tim Polzehl, André Beyer, Tibor Pölz, Sebastian Möller; T-Labs, Germany  
Show&Tell-1-8

This paper introduces a novel crowdsourcing platform provided to the community. The platform operates on mobile devices and makes data generation and labeling scenarios available for many related research tracks potentially covering also small and under-represented languages. Besides the versatile ways for commencing studies using the platform, also active research on crowdsourcing itself becomes feasible. With special focus on speech- and video recordings, the mobility and scalability of the platform is expected to stimulate and foster data-driven studies and insights throughout the community.

**On the Use of the ‘Pure Data’ Programming Language for Teaching and Public Outreach in Speech Processing**

Roger K. Moore; University of Sheffield, UK  
Show&Tell-1-9

Many educational institutions include a course on speech processing as part of their regular curriculum. Such courses usually cover basic principles in acoustics, phonetics and speech signal processing, and they often incorporate practical lab classes using specialised speech processing toolboxes such as ‘Praat’. Such toolkits are valuable resources for teaching and learning, but they often involve scripting solutions based on prescribed templates and non-real-time processing, and they do not lend themselves to be used by non-specialists, such as members of the general public or schoolchildren. This paper introduces the Pure Data open-source real-time visual programming language, and presents examples of its use for teaching and public outreach in speech processing. The ‘Show & Tell’ session will present ‘live’ examples as well as hands-on interactive demonstrations to illustrate its value.

**SyncWords: A Platform for Semi-Automated Closed Captioning and Subtitles**

Aleksandr Dubinsky; SyncWords, USA  
Show&Tell-1-10

As progressing social mores and new legislation mandate greater levels of accessibility across broadcast television, online video, education, and the workplace, it becomes ever more important to deliver captioned content at lower cost, with higher quality, and more quickly. To reach those goals, attention is shifting toward automation. While many attempts have been made to apply large vocabulary ASR transcription for this task, the results on noisy, post-processed, multi-speaker recordings are insufficient for most real-world applications, to say the least. More viable is the automated synchronization of a human-prepared transcript with the media, and automating the breaking up of the text into well-structured, meaningful captions. SyncWords is a platform that incorporates innovative, originally developed algorithms for both of these tasks, and presents an easy-to-use commercial Web interface for submitting work and receiving captions. In addition, an interactive review tool is available in which user corrections are fed back and cause a re-evaluation of the results in real time.

**Simple4All**

Robert A.J. Clark; Simple4All Project Consortium  
Show&Tell-1-11

The Simple4All project enables the building of speech synthesis systems automatically from data with little or no expert supervision, thus creating speech synthesisers for new languages or domains in a quick, easy and pain-free way. Whilst substantial effort has been invested over the years into developing novel and progressively more natural sounding speech synthesis techniques, there has always been a high entry barrier to building a speech synthesis system in a new language, because conventional methods require expensive linguistic resources, detailed expert knowledge,
and a high degree of skill in system ‘tuning’. The result is that only a very few of the world’s languages have speech synthesis systems. For those less widely-spoken languages that do have a system available, the quality of the synthetic speech is not very good, because of the lack of resources available and the lack of skills to ‘tune’ the system.

The main objectives of Simple4All are to address this imbalance by providing Open Source tools to:

- enable the creation of a voice for any domain in any language,
- provide an end-to-end framework in which these voices can be automatically built and improved,
- produce natural and expressive speech synthesis,
- provide feedback-driven learning to improve systems.

Keynote 4: William S.-Y. Wang
Garnet 213-218
08:30 – 09:30, Wednesday 17 September 2014
Chair: Chiu-yu Tseng

Sound Patterns in Language
William S.-Y. Wang; Chinese University of Hong Kong, China

In contrast to other species, humans are unique in having developed thousands of diverse languages which are not mutually intelligible. However, any infant can learn any language with ease, because all languages are based upon common biological infrastructures of sensori-motor, memorial, and cognitive faculties. While languages may differ significantly in the sounds they use, the overall organization is largely the same. It is divided into a discrete segmental system for building words and a continuous prosodic system for expressing, phrasing, attitudes, and emotions. Within this organization, I will discuss a class of languages called ‘tone languages’, which makes special use of F0 to build words. Although the best known of these is Chinese, tone languages are found in many parts of the world, and operate on different principles. I will also comment on relations between sound patterns in language and sound patterns in music, the two worlds of sound universal to our species.

Oral Session 21: Statistical Parametric Speech Synthesis
Garnet 213-218
10:00 – 12:00, Wednesday 17 September 2014
Chairs: Keiichi Tokuda and Rob Clark

Measuring the Perceptual Effects of Modelling Assumptions in Speech Synthesis Using Stimuli Constructed from Repeated Natural Speech
Gustav Eje Henter¹, Thomas Merritt¹, Matt Shannon², Catherine Mayo¹, Simon King¹; ¹University of Edinburgh, UK; ²University of Cambridge, UK

Acoustic models used for statistical parametric speech synthesis typically incorporate many modelling assumptions. It is an open question to what extent these assumptions limit the naturalness of synthesised speech. To investigate this question, we recorded a speech corpus where each prompt was read aloud multiple times. By combining speech parameter trajectories extracted from different repetitions, we were able to quantify the perceptual effects of certain commonly used modelling assumptions. Subjective listening tests show that taking the source and filter parameters to be conditionally independent, or using diagonal covariance matrices, significantly limits the naturalness that can be achieved. Our experimental results also demonstrate the shortcomings of mean-based parameter generation.
This paper presents an investigation of the separate perceptual degradations introduced by the modelling of source and filter features in statistical parametric speech synthesis. This is achieved using stimuli in which various permutations of natural, vocoded and modelled source and filter are combined, optionally with the addition of filter modifications (e.g. global variance or modulation spectrum scaling). We also examine the assumption of independence between source and filter parameters. Two complementary perceptual testing paradigms are adopted. In the first, we ask listeners to perform "same or different quality" judgements between pairs of stimuli from different configurations. In the second, we ask listeners to give an opinion score for individual stimuli. Combining the findings from these tests, we draw some conclusions regarding the relative contributions of source and filter to the currently rather limited naturalness of statistical parametric synthetic speech, and test whether current independence assumptions are justified.

Voice Expression Conversion with Factorised HMM-TTS Models

Javier Latorre, Vincent Wan, Kayoko Yanagisawa; Toshiba Research Europe, UK

This paper proposes a method to modify the expression or emotion in a sample of speech without altering the speaker’s identity. The method exploits a statistical speech model that factorises the speaker identity from expressions using linear transforms. For this approach, the set of transforms that best fit the speaker and expression of the input speech sample are learned. They are then combined with the expression transforms of the desired expression taken from another speaker. Since the combined expression transform is factorised and contains information about expression only, it may be applied to the original speech sample to modify its expression to the desired one without altering the identity of the speaker. Notably, this method may be applied universally to any voice without the need for a parallel training corpus.

Noise-Robust TTS Speaker Adaptation with Statistics Smoothing

Kayoko Yanagisawa, Langzhou Chen, Mark J.F. Gales; Toshiba Research Europe, UK

In practical scenarios for speaker adaptation of speech synthesis systems, the quality of adaptation audio data may be poor. In these situations, it is necessary to make use of the available audio to capture the speaker attributes, whilst aiming to obtain a synthetic voice which does not have any of the low-quality attributes of the audio. One approach to achieving this is to define a sub-space of parametric synthesis parameters in which the adapted system must lie. Though this yields reasonable synthesis quality, target speaker similarity degrades. Quality is also affected in severe noise conditions. This paper describes a smoothing approach that addresses this problem. For a noisy target speaker, first a 'similar speaker' is selected from a database of speakers. Statistics from this speaker are then smoothed with those obtained from the target speaker. By appropriately combining the two sources of information, it is possible to balance similarity and quality. Results indicate that both the quality and similarity can be improved by smoothing, especially for severe noise conditions. The similarity performance, however, varies from speaker to speaker, indicating the importance of a reasonable automatic speaker selection method and the coverage of the candidate speaker pool.

Speech Synthesis in Various Communicative Situations: Impact of Pronunciation Variations

Sandrine Brognaux 1, Benjamin Picart 2, Thomas Drugman 2; 1Université catholique de Louvain, Belgium; 2Université de Mons, Belgium

While current research in speech synthesis focuses on the generation of various speaking styles or emotions, very few studies have addressed the possibility of including phonetic variations according to the communicative situation of the target speech (sports commentaries, TV news, etc.). However, significant phonetic variations have been observed, depending on various communicative factors (e.g. spontaneous/read and media broadcast or not). This study analyses whether these alternative phonetic variations contribute to the plausibility of the message and should therefore be considered in synthesis. To this end, subjective tests are performed on synthesized French sports commentaries. They aim at comparing HMM-based speech synthesis with genuine pronunciation and with neutral NLP-produced phonetization. Results show that the integration of the phonetic variations significantly improves the perceived naturalness of the generated speech. They also highlight the relative importance of the various types of variations and show that schwa elisions, in particular, play a crucial role in that respect.

Formant-Controlled Speech Synthesis Using Hidden Trajectory Model

Ming-Qi Cai, Zhen-Hua Ling, Li-Rong Dai; USTC, China

This paper presents a statistical parametric speech synthesis method using hidden trajectory model (HTM) for flexibly controlling the formant positions and bandwidths of synthetic speech. In an HTM, hidden formant trajectories are generated by a bidirectional filtering process on the time-aligned and phone-dependent formant targets. The observed cepstral features are constituted by a formant-related component, which is predicted from the hidden formant trajectories using a nonlinear and analytical function, and a residual component, which is modeled by context-dependent Gaussians. In this paper, we apply HTM-based acoustic modeling to speech synthesis. The distribution parameters of the formant targets are manipulated at synthesis time to control the characteristics of synthetic speech. In our implementation, the distributions of residual cepstra are estimated for each quinphone and the question set used in the decision-tree-based model clustering is tailored so as to acquire high controllability for vowels. Experimental results show that this proposed method can achieve effective controllability on the formant positions and bandwidths while keeping almost the same naturalness as the conventional HMM-based approach.

NOTES
Boosted Deep Neural Networks and Multi-Resolution Cochleagram Features for Voice Activity Detection
Xiao-Lei Zhang\textsuperscript{1}, DeLiang Wang\textsuperscript{2}, Tsinghua University, China; \textsuperscript{2}Ohio State University, USA

Voice activity detection (VAD) is an important front-end of many speech processing systems. In this paper, we describe a new VAD algorithm based on boosted deep neural networks (bDNNs). The proposed algorithm first generates multiple base predictions for a single frame from only one DNN and then aggregates the base predictions for a better prediction of the frame. Moreover, we use a new acoustic feature, multi-resolution cochleagram (MRCG), that concatenates the cochleagram features at multiple spectrotemporal resolutions and shows superior speech separation results over many acoustic features. Experimental results show that bDNN-based VAD with the MRCG feature outperforms state-of-the-art VADs by a considerable margin.

Towards Improving Statistical Model Based Voice Activity Detection
Ming Tu, Xiang Xie, Yishan Jiao; BIT, China

Statistical model based voice activity detection (VAD) is commonly used in various speech related research and applications. In this paper, we try to improve the performance of statistical model based VAD via a new feature extraction method. Our main innovation focuses on that we apply Mel-frequency subband coefficients with power-law nonlinearity as feature for statistical model based VAD instead of Discrete Fourier Transform (DFT) coefficients. This proposed feature is then modeled by Gaussian distribution. Performances of this method are comprehensively compared with existing methods. Meanwhile we also test power-law nonlinearity on existing methods. Experimental results prove that with proposed subband coefficients the performance of statistical model based VAD could be improved a lot. Power-law nonlinearity on DFT coefficients could also bring some improvement.

The Use of Low-Frequency Ultrasound for Voice Activity Detection
Ian Vince McLoughlin; USTC, China

An active detection system is developed which uses low-power low-frequency ultrasonic reflection to determine the lip state (i.e. whether open, closed or in between) of a human speaker and hence the presence of vocal activity. In operation, a small loudspeaker or sounder, located within a few centimetres of the lips, produces an excitation signal which is emitted towards the lips. A co-located microphone receives the signal reflected from the lip region. Even simple analysis of the reflected information reveals whether the mouth is open, closed or in between. Careful placement of the excitation signal at the extreme low end of the ultrasonic range, allows its generation and analysis to be done with inexpensive off-the-shelf audio hardware. This paper describes the techniques used, presents experimental details regarding the signals, then implement and evaluates a simple voice activity detector based on the technique.

Notes

Unlike traditional speech corpora, the audio recordings in Apollo are extensive from a longitudinal perspective (i.e., 6–12 days each). From SAD perspective, the data offers many challenges: (i) noise distortion with variable SNR, (ii) channel distortion, and (iii) extended periods of non-speech activity. Here, we use the recently proposed Combo-SAD, which has performed remarkably well in DARPA RATS evaluations, as our baseline system [2]. Our analysis reveals that the Combo-SAD performs well when speech-pause durations are balanced in the audio segment, but deteriorates significantly when speech is sparse or absent. In order to mitigate this problem, we propose a simple yet efficient technique which builds an alternative model of speech using data from a separate corpora, and embeds this new information within the Combo-SAD framework. Our experiments show that the proposed approach has a major impact on SAD performance (i.e., +30% absolute), especially in audio segments that contain sparse or no speech information.
Improving the Speech Activity Detection for the DARPA RATS Phase-3 Evaluation

Jeff Ma; Raytheon BBN Technologies, USA

Wed 022-6

This paper presents the work that we conducted for building the speech activity detection (SAD) systems for the phase 3 evaluation of the RATS program. The work focused on improving the SAD performance with the neural network (NN) approach. The major efforts include reducing the false rejections errors by extensions of speech regions in the training references and use of post-processing NNs, and removing channel variations by design of channel bottleneck features with the deep NN learning approach. With these efforts more 25% relative improvements were achieved over the phase 2 evaluation system. The bigger contribution of the design of the bottleneck features was the enhancement of the SAD system performance on new channels. Our results revealed that the bottleneck features were able to improve SAD performance on new channels significantly.

Oral Session 23: Disordered Speech

Peridot 204-205
10:00–12:00, Wednesday 17 September 2014
Chairs: Emily Mower Provost and Jordan Green

Modeling Pronunciation, Rhythm, and Intonation for Automatic Assessment of Speech Quality in Aphasia Rehabilitation

Duc Le, Emily Mower Provost; University of Michigan, USA

Wed 023-1

Patients with aphasia often have impaired speech-language production skills, resulting in tremendous difficulties in tasks that require verbal communication. To facilitate rehabilitation outside of therapy, we are collaborating with the University of Michigan Aphasia Program (UMAP) to develop an automated system capable of providing feedback regarding the patient’s verbal output. In this paper we introduce a robust method for extracting rhythm and intonation features from aphasic speech based on template matching. These features, combined with Goodness of Pronunciation (GOP) scores and our previous feature set, help our system achieve human-level performance in classifying the quality of speech produced by patients attending UMAP. The results presented in this work demonstrate the efficacy of our technique and the potential of this system for handling natural speech data recorded in non-ideal conditions as well as the unpredictability in aphasic speech patterns.

Ranking Severity of Speech Errors by their Phonological Impact in Context

Sofia Strömbergsson 1, Christina Tännander 2, Jens Edlund 1, KTH, Sweden; 2MTM, Sweden

Wed 023-2

Children with speech disorders often present with systematic speech error patterns. In clinical assessments of speech disorders, evaluating the severity of the disorder is central. Current measures of severity have limited sensitivity to factors like the frequency of the target sounds in the child’s language and the degree of phonological diversity, which are factors that can be assumed to affect intelligibility. By constructing phonological filters to simulate eight speech error patterns often observed in children, and applying these filters to a phonologically transcribed corpus of 350K words, this study explores three quantitative measures of phonological impact: Percentage of Consonants Correct (PCC), edit distance, and degree of homonymy. These measures were related to estimated ratings of severity collected from 34 practicing clinicians. The results show an expected high correlation between the PCC and edit distance metrics, but that none of the three metrics align with clinicians’ ratings. Although these results do not generate definite answers to what phonological factors contribute the most to (un)intelligibility, this study demonstrates a methodology that allows for large-scale investigations of the interplay between phonological errors and their impact on speech in context, within and across languages.

Automatic Detection of Parkinson’s Disease from Words Uttered in Three Different Languages

J.R. Orozco-Arroyave 1, Florian Högni 2, J.D. Arias-Londoño 1, J.F. Vargas-Bonilla 1, S. Skodda 2, J. Rusz 4, Elmar Nöth 2; 1Universidad de Antioquia, Colombia; 2FAU Erlangen-Nürnberg, Germany; 3Ruhr-Universität Bochum, Germany; 4Czech Technical University in Prague, Czech Republic

Wed 023-3

About 90% of the people with Parkinson’s disease (PD) develop speech impairments such as monopitch, monoloudness, imprecise articulation, and other symptoms. There are several studies addressing the problem of the automatic detection of PD from speech signals in order to develop computer aided tools for the assessment and monitoring of the patients. Recent works have shown that it is possible to detect PD from speech with accuracies above 90%; however, it is still unclear whether it is possible to make the detection independent of the spoken language. This paper addresses the automatic detection of PD considering speech recordings of three languages: German, Spanish and Czech. According to the results it is possible to classify between speech of people with PD and healthy controls (HC) with accuracies ranging from 84% to 99%, depending on the utterance.

Automating an Objective Measure of Pediatric Speech Intelligibility

Jason Lilley 1, Susan Nittrouer 2, H. Timothy Bunnell 1; 1Nemours Biomedical Research, USA; 2Ohio State University, USA

Wed 023-4

There are many research and clinical settings in which accurate objective measures of speech intelligibility are either necessary or highly desirable. Unfortunately, most objective measures of intelligibility require blinded human listeners to classify speech tokens obtained from the talker whose speech intelligibility is to be measured, a process that can be both costly and time-consuming to carry out. In a research setting, it is often possible to justify the time and cost required, but in a clinical setting this is usually infeasible. The current work describes an effort to develop tools that could be of practical use in a clinical setting. An automatic response scoring system based on 32-Gaussian mixture hidden Markov models was trained on responses of both children with normal hearing and those with hearing loss to the Children’s Speech Intelligibility Measure (CSIM) [1]. The system was able to predict a human listener’s response out of 12 choices over 60% of the time. Aggregate CSIM scores computed from the ASR system had a high correlation with scores compiled by human listeners (r2=.83). Automatic scoring of utterances from children could allow the CSIM and similar testing procedures to be used more frequently in research and clinical settings.

NOTES
Acoustic limitations that parallel deficits in speech intelligibility. vocal tract, but that dysarthric talkers may nonetheless exhibit idiosyncratic movement strategies in learning to control a virtual vocal tract. Results show that the hybrid DNN-HMM outperforms the conventional GMM-HMM for all experiments on both normal and disordered speech. The total correctness accuracy of the system at the phoneme level is above 85% when used with disordered speech.

Acoustic and Kinematic Characteristics of Vowel Production Through a Virtual Vocal Tract in Dysarthria

Jeff Berry, Andrew Kolb, Cassandra North, Michael T. Johnson; Marquette University, USA

Broadening our understanding of the components and processes of speech sensorimotor learning is crucial to furthering methods of speech neurorehabilitation. Recent research in limb sensorimotor control has used virtual environments to study learning in novel sensorimotor working spaces. Comparable experimental paradigms have yet to be undertaken in speech learning. We present acoustic and kinematic data obtained from participants producing vowels in unfamiliar articulatory-acoustic conditions. We compare between two different acoustic models, the conventional GMM-HMM and the hybrid DNN-HMM. Results show that the hybrid DNN-HMM outperforms the conventional GMM-HMM for all working spaces using a virtual vocal tract. Talkers with dysarthria and healthy controls were asked to produce vowels using an electromagnetic articulograph-driven speech synthesizer for participant-controlled auditory feedback. The aim of the work was to characterize performance within and between groups to generate hypotheses regarding experimental manipulations that may bolster our understanding of speech sensorimotor learning. Results indicate that dysarthric talkers displayed relatively reduced acoustic working spaces and somewhat more variable acoustic targets compared to controls. Kinematic measures of articulatory dynamics, particularly peak speed and movement jerk-cost, were idiosyncratic and did not dissociate talker groups. These findings suggest that individuals with dysarthria and healthy talkers may use idiosyncratic movement strategies in learning to control a virtual vocal tract, but that dysarthric talkers may nonetheless exhibit acoustic limitations that parallel deficits in speech intelligibility.

This paper introduces a pronunciation verification method to be used in an automatic assessment therapy tool of child disordered speech. The proposed method creates a phone-based search lattice that is flexible enough to cover all probable mispronunciations. This allows us to verify the correctness of the pronunciation and detect the incorrect phonemes produced by the child. We compare between two different acoustic models, the conventional GMM-HMM and the hybrid DNN-HMM. Results show that the hybrid DNN-HMM outperforms the conventional GMM-HMM for all experiments on both normal and disordered speech. The total correctness accuracy of the system at the phoneme level is above 85% when used with disordered speech.

This article gives an overview of the EMG-UKA corpus, a corpus of electromyographic (EMG) recordings of articulatory activity enabling speech processing (in particular speech recognition and synthesis) based on EMG signals, with the purpose of building Silent Speech interfaces. Data is available in multiple speaking modes, namely audibly spoken, whispered, and silently articulated speech. Besides the EMG data, synchronous acoustic data was additionally recorded to serve as a reference. The corpus comprises 63 recorded sessions from 8 speakers, the total amount of data is 7-32 hours. A trial subset consisting of 1-32 hours of data, is freely available for download.

Whispered speech is a natural mode of speech in which voicing is absent — its acoustics differ significantly from normally spoken speech or so-called neutral speech, such that it is challenging to use only neutral speech to build speech processing and automatic recognition systems that can deal effectively with whispered speech. The same time, humans can naturally produce and perceive whispered speech without explicit training. Tonal languages such as Mandarin present an interesting dilemma — tone is primarily encoded by pitch tracks which are absent during whispered speech, but humans can still tell tones apart. How humans manage to process whispered speech well without explicit training on it, whereas machine algorithms fail, is presently an unresolved question which could prove fruitful with study. This, however, is hindered by the lack of suitable, systematically collected corpora. We present iWhisper-Mandarin, a 25-hour parallel corpus of neutral and whispered Mandarin, designed to support research in linguistics and speech technology. We demonstrate and verify that earlier techniques applied to whispered speech from non-tonal languages also work with Mandarin, and present some preliminary studies on voice activity detection and whispered Mandarin speech recognition.

Euronews: A Multilingual Benchmark for ASR and LID

Roberto Gretter; FBK, Italy

In this paper we present the first recognition experiments on a multilingual speech corpus, designed for Automatic Speech Recognition (ASR) and Language IDentification (LID) purposes. Data come from the portal Euronews and were acquired both from the Web and from TV. The corpus includes data in 10 languages.
In this paper we present an overview on why speech phase spectrum is not so widely spread. In this paper, we present an overview on why speech phase spectrum has been neglected in the conventional techniques used in different applications including: speech separation/enhancement, automatic speech and speaker recognition and speech synthesis. We proceed with giving highlights on the recent progress carried out in demonstrating the importance of phase in different applications and how it impacts on the overall performance. The paper is an introduction to the Interspeech 2014 special session phase importance in speech processing applications.

Phase-Based Harmonic/Percussive Separation
Estefania Cano 1, Mark Plumbley 2, Christian Dittmar 1; 1Fraunhofer IDMT, Germany; 2Queen Mary University of London, UK

In this paper, a method for separation of harmonic and percussive elements in music recordings is presented. The proposed method is based on a simple spectral peak detection step followed by a phase expectation analysis that discriminates between harmonic and percussive components. The proposed method was tested on a database of 10 audio tracks and has shown superior results to the reference state-of-the-art approach.
Phase Distortion Statistics as a Representation of the Glottal Source: Application to the Classification of Voice Qualities
Gilles Degottex$^{1,}$, Nicolas Obin$^{2,}$; $^{1}$University of Crete, Greece; $^{2}$IRCAM, France

The representation of the glottal source is of paramount importance for describing para-linguistic information carried through the voice quality (e.g., emotions, mood, attitude). However, some existing representations of the glottal source are based on analytical glottal models, which assume strong a priori constraints on the shape of the glottal pulses. Thus, these representations are restricted to limited number of voices. Recent progresses in the estimation of the glottal models revealed that the Phase Distortion (PD) of the signal carries most of the information about the glottal pulses. This paper introduces a flexible representation of the glottal source - based on the short-term modelling of the phase distortion. This representation is not constrained by a specific analytical model, and thus can be used to describe a larger variety of expressive voices. We address the efficiency of this representation for the recognition of various voice qualities, with comparison to MFCC and standard glottal source representations.

A Measure of Phase Randomness for the Harmonic Model in Speech Synthesis
Gilles Degottex$^{1,}$, Daniel Erro$^{2,}$; $^{1}$University of Crete, Greece; $^{2}$Universidad del País Vasco, Spain

Modern statistical speech processing frameworks require the speech signals to be translated into feature vectors by means of vocoders. While features representing the amplitude envelope already exist (e.g. MFCC, LSF), parametrizing the phase information is far from straightforward, not only because it is a circular data, but also because it shows an irregular behaviour in noisy time-frequency regions. Thus, many vocoders reconstruct speech by using minimum phases and random phases, relying on a previous voicing decision. In this paper, a phase feature is suggested to represent the randomness of the phase across the full time-frequency plan, in both voiced and unvoiced segments, without voicing decision. Resynthesis experiments show that, when integrated into a full-band harmonic vocoder, the suggested randomization feature is slightly better, on average, to STRAIGHT’s aperiodicity. In HMM-based synthesis, the results show that the suggested vocoder reduces the complexity of the analysis and statistical modelling by removing the voicing decision, while keeping the perceived quality.

Enhancement of Speech Intelligibility in Near-End Noise Conditions with Phase Modification
Emma Jokinen$^{1,}$, Marko Takanen$^{1,}$, Hannu Palakka$^{2,}$, Paavo Alku$^{1,}$; $^{1}$Aalto University, Finland; $^{2}$Microsoft, Finland

Post-processing methods can be used in mobile communications to improve the intelligibility of speech in adverse near-end noise conditions. This study proposes a phase spectrum modification method for intelligibility enhancement in mobile devices. The method first modifies the phase spectrum of speech in order to reduce the amplitude range of the signal. The amplitude level is then equalized to that of the unprocessed speech hence resulting in amplification of signal energy. The performance of the proposed method was evaluated in comparison to a known post-processing method with an objective intelligibility metric in multiple noise conditions. Results indicate that the proposed post-processing method improves speech intelligibility over the reference method.

A Hybrid Approach to Segmentation of Speech Using Group Delay Processing and HMM Based Embedded Reestimation
S. Aswin Shanmugam, Hema Murthy; IIT Madras, India

The most popular method for automatic segmentation is embedded reestimation of monophone HMMs after flat start initialization, followed by forced alignment. This method may not yield accurate boundaries. To address this issue, group delay based processing of short-time energy (STE) is performed on the speech signal to obtain syllable boundaries. The syllable boundaries are accurate, but there are a number of spurious insertions as the text transcription is not used during segmentation. The boundaries obtained using group delay segmentation in the vicinity of the HMM syllable boundaries are used as correct boundaries to reestimate the monophone HMM models, where the monophone HMMs are restricted to the syllable boundaries rather than the whole utterance. The reestimated boundaries are again compared with the group delay boundaries and corrected again. Essential signal processing for detecting boundaries and statistical segmentation for acoustic modelling work in tandem to obtain accurate segmentation at both phoneme and syllable levels. Considering phones and syllables as basic units, HMM based speech synthesis systems (HTS) are built with the proposed segmentation method. Listening tests indicate that there is an improvement in the quality of synthesis.

The Importance of Phase on Voice Quality Assessment
Maria Koutsogiannaki, Olympia Simantiraki, Gilles Degottex, Yannis Stylianou; University of Crete, Greece

State of the art objective measures for quantifying voice quality mostly consider estimation of features extracted from the magnitude spectrum. Assuming that speech is obtained by exciting a minimum-phase (vocal tract filter) and a maximum-phase component (glottal source), the amplitude spectrum cannot capture the maximum phase characteristics. Since voice quality is connected to the glottal source, the extracted features should be linked with the maximum-phase component of speech. This work proposes a new metric based on the phase spectrum for characterizing the maximum-phase component of the glottal source. The proposed feature, the Phase Distortion Deviation, reveals the irregularities of the glottal pulses and therefore, can be used for detecting voice disorders. This is evaluated in a ranking problem of speakers with spasmodic dysphonia. Results show that the obtained ranking is highly correlated with the subjective ranking provided by doctors in terms of overall severity, tremor and jitter. The high correlation of the suggested feature with different metrics reveals its ability to capture voice irregularities and highlights the importance of the phase spectrum in voice quality assessment.

Feature Extraction from Analytic Phase of Speech Signals for Speaker Verification
Karthika Vijayan, Vinay Kumar, K. Sri Rama Murty; IIT Hyderabad, India

The objective of this work is to study the speaker-specific nature of analytic phase of speech signals. Since computation of analytic phase suffers from phase wrapping problem, we have used its derivative- the instantaneous frequency for feature extraction.
The cepstral coefficients extracted from smoothed subband instantaneous frequencies (IFCC) are used as features for speaker verification. The performance of IFCC features is evaluated on NIST-2003 speaker recognition evaluation database and is compared with baseline mel-frequency cepstral coefficients (MFCC). The performance of IFCC features is observed to be comparable with MFCC features in terms of equal error rates and minimum detection cost function values. Different strategies for evaluating the speaker verification performance of IFCC and MFCC are considered and it is found that the evaluation based on cosine similarity delivers better performance than other strategies under consideration.

A Cross-Vocoder Study of Speaker Independent Synthetic Speech Detection Using Phase Information
Jon Sanchez, Ibon Saratxaga, Inma Hernandez, Eva Navas, Daniel Erro; Universidad del Pais Vasco, Spain

Current speaker verification systems are vulnerable to advanced speech manipulation techniques such as voice conversion and speaker adaptation for TTS systems. Effective anti-spoofing systems that allow the discrimination between human and synthetic impostors have been developed. However, many of them still present two main drawbacks: speaker dependency and, more importantly, counterfeiting technique dependency. Thus, getting a universal synthetic speech detector that performs better than MFCC for multi-vocoder models.

Adapting Prosodic Chunking Algorithm and Synthesis System to Specific Style: The Case of Dictation
Elisabeth Delais-Roussarie 1, Damien Lolive 2, Hyon Yoo 1, Nelly Barbot 2, Olivier Rosec 3, 1 LLF (UMR 7110), France; 2 IRISA, France; 3 Voxgen, France

In this paper, we present an approach that allows a TTS-system to dictate texts to primary school pupils, while being in conformity with the prosodic features of this speaking style. The approach relies on the elaboration of a preprocessing prosodic module that avoids developing a specific system for a so limited task. The proposal is based on two distinct elements: (i) the results of a preliminary evaluation that allowed getting feedback from potential users; (ii) a corpus study of 10 dictations annotated or uttered by 13 teachers or speech therapists (10 and 3 respectively).

The preliminary evaluation focused on three points: the accuracy of the segmentation procedure, the size of the automatically calculated chunks, and the intelligibility of the synthesized voice. It showed that the chunks were judged too long, and the speaking rate too fast. We thus decided to work on these two issues while analyzing the collected data, and confronting the obtained realizations with the outcome of the speech synthesis system and the chunking algorithm. The results of the analysis lead to propose a module that provides for this speaking style an enriched text that can be treated by the synthesizer to constrain the unit selection and the prosodic realization.

The Articulation of Lexical and Post-Lexical Palatalization in Korean
Jae-Hyun Sung; University of Arizona, USA

Palatalization in Korean is of two types — lexical palatalization governed by language-specific phonological rules, and post-lexical palatalization that appears to be purely phonetic. While lexical palatalization only occurs when a morpheme boundary intervenes between a target consonant and a palatalization trigger, post-lexical palatalization occurs irrespective of the presence of a morpheme boundary. This study investigates whether these two types of palatalization and different morphological structures of words manifest as distinct tongue gestures using ultrasound imaging of 4 native speakers of Korean. Comparison of the ultrasound tongue contours shows that the gestural distinction between lexical and post-lexical palatalization may not be the same across individual speakers. Furthermore, the effects of morpheme boundaries are not uniform across different coronal consonants and speakers in terms of tongue gestures. The findings from this study provide further empirical evidence for the role of morphological structures in coarticulation, and are in line with mounting evidence for speaker-specific variability in speech production.

NOTES
Articulation and Neutralization: A Preliminary Study of Lenition in Scottish Gaelic
Diana Archangeli 1, Samuel Johnston 2, Jae-Hyun Sung 2, Muriel Fisher 2, Michael Hammond 2, Andrew Carnegie 2; 1 University of Hong Kong, China; 2 University of Arizona, USA

Wed-P-16-4, Poster

Initial Consonant Mutation in Scottish Gaelic is considered to be morphological, somewhat idiosyncratic, and neutralizing, that is, merging either the mutated sound and some underlying sound or merging two mutated sounds. This study explores articulation in one class of mutation, called Lenition (also Aspiration), asking the question: whether these sounds are articulated in the same fashion or not. Comparison of relevant ultrasound images collected from 3 native speakers of Scottish Gaelic shows that speakers maintain distinctions between True Lenition and False Lenition, suggesting that there is incomplete neutralization. Furthermore, when Lenition of two distinct sounds converge on the same target, a neutralization effect is observed. The results are consistent with a phonological model which distinguishes between surface forms corresponding to different sources, showing very little complete articulatory neutralisation.

Nasality in Speech and its Contribution to Speaker Individuality
Kanae Amino 1, Hisanori Makinae 1, Tatsuya Kitamura 2; 1 NRIPS, Japan; 2 Konan University, Japan

Wed-P-16-5, Poster

The term nasality refers to the timbre of the nasal phonemes. It is also used to express the quality of sound that characterises some speakers. In this paper, we propose to classify nasality in natural speech into four types: phonemic nasality, nasality in assimilation, incidental nasality in the production of voiced plosives, and nasality associated with speaker individuality. Speech sounds recorded separately for oral and nasal outputs were analysed and the four types of nasality were observed individually. In order to investigate the relationship between the nasality in running speech and the perception of speaker similarity, we conducted an experiment. The results revealed that listeners rated speaker similarity exploiting phonemic nasality when it existed in the utterance and also used speaker-related nasality regardless of the existence of phonemic nasals.

Is Speech Rhythm an Intrinsic Property of Language?
Jason Brown, Eden Matene; University of Auckland, New Zealand

Wed-P-16-6, Poster

Different languages have traditionally been classified into different rhythm types. Most studies of rhythm have either implicitly or explicitly accepted that rhythm is an inherent property of a language. This study aims to determine whether rhythm is an intrinsic property of languages, or whether rhythm is an epiphenomenal byproduct of the phonotactic structures of a given stimulus. The question that this project addresses is to what extent the phonological properties of a language can be correlated with rhythmic categories; for instance, whether a language has consonant clusters, makes use of contrastive tone, has complex syllables, exhibits vowel reduction, etc. and whether these can be linked to what kind of rhythmic profile a language fits into.

Where /aR/ is the /R/ in Standard Austrian German?
Anke Jackschna 1, Barbara Schuppler 2, Rudolf Muhr 1; 1 Karl-Franzens-Universität Graz, Austria; 2 Technische Universität Graz, Austria

Wed-P-16-7, Poster

The present paper investigates the conditions under which different realizations of /R/ occur in standard Austrian German. The study is based on 309 word tokens containing the phone sequence /aR/ in coda position drawn from a corpus of read speech from seven male Austrian radio speakers. Acoustic measurements of the vowel /a/ revealed that F1, F2 and F3 are significant predictors for the realization of /R/ as either trill, fricative or as absent. Moreover, /a/ tends to be longer when /R/ is absent than when it is present. Our analysis of the linguistic conditions for the different realizations of /R/ showed that /R/ is least reduced in stressed syllables and in words read in isolation. Furthermore, we observe that the segmental context significantly affects the realization of /R/. Most importantly, we find significant effects of morphology: /R/ tends to be more reduced when it is part of a grammatical morpheme than when it is part of the stem of a word. These findings inform the further development of models of pronunciation variation for human and automatic speech recognition.

Diphthongized Vowels in the Yi County Hui Chinese Dialect
Fang Hu, Minghui Zhang, Chinese Academy of Social Sciences, China

Wed-P-16-8, Poster

This paper is an acoustic phonetic description of the vowels and diphthongs in the Yi county dialect and compares the diphthongized vowels with monophthongs and diphthongs in terms of their temporal structures and spectral characteristics. Results show that rising diphthongs have two targets while falling diphthongs only have one dynamic target. Diphthongized vowels occur as an intermediate category between monophthongs and diphthongs. Diphthongized vowels have distinctive onsets but a neutralized offset.

Rhythmic Variability Between Some Asian Languages: Results from an Automatic Analysis of Temporal Characteristics
Volker Dellwo 1, Peggy Mok 2, Mathias Jenny 1; 1 Universität Zürich, Switzerland; 2 Chinese University of Hong Kong, China

Wed-P-16-9, Poster

The rhythmic organization of speech can vary between languages. In the present research we studied rhythmic variability between Mandarin, Cantonese and Thai using automatically retrieved prosodic temporal characteristics from read speech. We measured the variability of intervals between amplitude peaks in the amplitude envelope (<10 Hz) and the durational characteristics of intervals with and without glottal activity (voiced and unvoiced intervals) in speech. Results for between language comparisons revealed significant differences between languages in both amplitude peak interval variability and voiced-voiceless interval durational characteristics. Results are discussed in connection with language specific phonotactic/phonological properties and hypotheses about the perceptual significance of the acoustic measurements in terms of speech rhythm.

NOTES
Listener Estimation of Speaker Age Based on Whispered Speech

Angelika Braun, Daniela Decker; Universität Trier, Germany

Wed-P-16-10, Poster

The assessment of a given speaker’s age has been shown to rest on laryngeal as well as vocal tract features. This study attempts to separate these two elements by comparing listener performance based on whispered and phonated speech. A total of 45 speakers belonging to three different age groups (20-35; 45-60; 70 and above) were assessed by 20 listeners. Results show that although listener performance decreases with whispered speech, it is still well above chance level. No significant difference was found between age groups, but female speakers were assessed correctly more often than males in both conditions. There are forensic implications to this finding, since in that context, age estimation sometimes has to be carried out on whispered speech, e.g. if whispering is used as a voice disguise.

The Lombard Effect with Thai Lexical Tones: An Acoustic Analysis of Articulatory Modifications in Noise

Benjawan Kasisopa, Virginie Attina, Denis Burnham; University of Western Sydney, Australia

Wed-P-16-11, Poster

When in a noisy environment speakers modify their speech production by increasing loudness, vowel duration, and fundamental frequency (F0), a phenomenon known as Lombard speech. Here Lombard speech in Thai was investigated in order to determine the effects of noise on the realisation of F0 in Thai lexical tones. Analysis of the acoustic characteristics of the five Bangkok Thai tones, in both continuous speech and citation form, produced in noise and in quiet showed that F0 is heightened in Lombard compared with clear speech. In addition, generally the contour of the tones also changes in Lombard speech; contours tend to be exaggerated towards the end of the tone in a direction consistent with the contour of the tone.

Intrinsic Spectral Analysis Based on Temporal Context Features for Query-by-Example Spoken Term Detection

Peng Yang 1, Cheung-Chi Leung 2, Lei Xie 1, Bin Ma 2, Haizhou Li 2; 1Northwestern Polytechnical University, China; 2A*STAR, Singapore

Wed-P-17-1, Poster

We investigate the use of intrinsic spectral analysis (ISA) for query-by-example spoken term detection (QbE-STD). In the task, spoken queries and test utterances in an audio archive are converted to ISA features, and dynamic time warping is applied to match the feature sequence in each query with those in test utterances. Motivated by manifold learning, ISA has been proposed to recover from untranscribed utterances a set of nonlinear basis functions for the speech manifold, and shown with improved phonetic separability and inherent speaker independence. Due to the coarticulation phenomenon in speech, we propose to use temporal context information to obtain the ISA features. Gaussian posteriorigram, as an efficient acoustic representation usually used in QbE-STD, is considered a baseline feature. Experimental results on the TIMIT speech corpus show that the ISA features can provide a relative 13.5% improvement in mean average precision over the baseline features, when the temporal context information is used.

Recent Improvements in SRI’s Keyword Detection System for Noisy Audio

Julien van Hout, Vikramjit Mittra, Yun Lei, Dimitra Vergyri, Martin Graciarena, Arindam Mandal, Horacio Franco; SRI International, USA

Wed-P-17-2, Poster

We present improvements to a keyword spotting (KWS) system that operates in highly adverse channel conditions with very low signal-to-noise ratio levels. We employ a system combination approach by combining the outputs of multiple large vocabulary continuous speech recognition (LVCSR) systems. These systems are complementary thanks to different design decisions across all levels of information: three speech activity detections systems; a wide range of front-end signal processing features (standard cepstral and filter-bank features, noise-robust features and multi-layer perceptron features); three statistical acoustic model types (Gaussian mixtures models, deep and convolutional neural networks); two keyword search strategies (word-based and phone-based). We explore the scenario where the keywords are known in advance by adding them to the language model and assigning higher weights to n-grams with keywords in them. The scores of each individual system are fused by a logistic-regression based classifier to produce the final system combination output. We present the performance of our system in the Phase III evaluations of DARPA’s Robust Automatic Transcription of Speech (RATS) program for Levantine Arabic and Farsi conversational speech corpora.

Utilizing State-Level Distance Vector Representation for Improved Spoken Term Detection by Text and Spoken Queries

Mitsuaki Makino, Naoki Yamamoto, Atsuhiko Kai; Shizuoka University, Japan

Wed-P-17-3, Poster

In spoken term detection (STD) systems, approximate subword-level matching of query term and automatically transcribed spoken documents is often employed for its reasonable accuracy and efficiency. However, high out-of-vocabulary (OOV) rate often degrades the subword-level recognition accuracy and affect the STD performance. This paper describes the usage of new expanded acoustic representations of subword sequence for improved scoring between OOV query term and subword-unit transcription. Each subword is expanded in corresponding subword’s HMM states and each state is represented as a new acoustic structural feature, a distribution-distance vector (DDV). The proposed DDV representation and scoring is easily combined with two typical baseline STD approaches: a DTW-based approximate matching with subword-level acoustic dissimilarity measure and a lattice-based confidence scoring of subword n-grams. The experimental result showed that the proposed DDV-based scoring method significantly outperforms the simple DTW-scoring baseline with very little increase in the required search time. The combination of the DDV-based scoring with the confidence-based scoring showed the complementary effect and attained the best STD performance compared with the NTCIR-10 SpokenDoc2(SDPWS) submitted results when only the NTCIR reference automatic transcript is used. A preliminary ex-

NOTES
experiment with spoken query terms also showed that the significant improvement for OOV queries.

Unsupervised Spoken Word Retrieval Using Gaussian-Bernoulli Restricted Boltzmann Machines
Raghavendra Reddy Pappagari, Shekhar Nayak, K. Sri Rama Murty; IIT Hyderabad, India

The objective of this work is to explore a novel unsupervised framework, using Restricted Boltzmann machines, for Spoken Word Retrieval (SWR). In the absence of labelled speech data, SWR is typically performed by matching sequence of feature vectors of query and test utterances using dynamic time warping (DTW). In such a scenario, performance of SWR system critically depends on representation of the speech signal. Typical features, like mel-frequency cepstral coefficients (MFCC), carry significant speaker-specific information, and hence may not be used directly in SWR system. To overcome this issue, we propose to capture the joint density of the acoustic space spanned by MFCCs using Gaussian-Bernoulli restricted Boltzmann machine (GBRBM). In this work, we have used hidden activations of the GBRBM as features for SWR system. Since the GBRBM is trained with speech data collected from large number of speakers, the hidden activations are more robust to the speaker-variability compared to MFCCs. The performance of the proposed features is evaluated on Telugu broadcast news data, and an absolute improvement of 12% was observed compared to MFCCs.

Unsupervised Query-by-Example Spoken Term Detection Using Bag of Acoustic Words and Non-Segmental Dynamic Time Warping
Basil George, Abhijeet Saxena, Gautam Mantena, Kishore Prahallad, B. Yegnaranayana; IIIT Hyderabad, India

The paper proposes an unsupervised framework to address the problem of spotting spoken terms in large speech databases. A two-stage retrieval mechanism is used to perform spoken term detection. A very efficient Bag of Acoustic Words (BoAW) index is created for quick retrieval of relevant documents. Using an N-gram approach, the optimum choice of acoustic dictionary that best describes the document is obtained. Once a quick reduction in search space is achieved in the first phase, the results are fed to the second stage of the retrieval engine. Here, a computationally optimised variant of dynamic programming, called Non-Segmental Dynamic Time Warping (NS-DTW), is used to further prune the results. All the experiments are conducted on MediaEval 2012 dataset. Performance is evaluated at the output of each stage, and the optimum parameters are obtained. We show that the cascade of these two stages helps in reducing the probable search space, which translates to higher search speeds, while ensuring comparable performance. The significance of the indexing framework is well in different scenarios and for different tasks, and there are very few diagnostic techniques for improving our understanding. In this paper, we present two diagnostic measurements that can be used to directly assess the quality of alignments between sequences of features independently of the intended use of the alignments downstream. We argue that such diagnostic techniques are valuable for intrinsically assessing speech features and alignment algorithms for keyword detection.

An Empirical Study of Multilingual and Low-Resource Spoken Term Detection Using Deep Neural Networks
Jie Li, Xiaorui Wang, Bo Xu; Chinese Academy of Sciences, China

As a further step of our previous work, this paper focuses on how to promote the multilingual spoken term detection (STD) system by the use of shared-hidden-layer multilingual DNN (SHL-MDNN). Seven languages namely Arabic, English, German, Japanese, Korean, Mandarin and Spanish are used in our experiments. Compared with our original multilingual STD system, which is based on Subspace GMMs (SGMMs), the resulting system reduces the average equal error rate (EER) on seven languages by 17.2%. Our STD system is also evaluated under low-resource conditions in this paper. We choose Mandarin and English as two target languages and simulate different degrees of available resources. The experimental results show that with the help of cross-lingual model transfer, our STD system can be elevated a lot in low-resource settings. To further improve the performance, we also attempt to use dropout strategy during the process of cross-lingual model transfer. However, no significant improvement can be observed in our experiments. This indicates the dropout method is not so effective on cross-lingual model transfer task.

Diagnostic Techniques for Spoken Keyword Discovery
Peter Schulum1, Murat Akbacak2; 1 Johns Hopkins University, USA; 2 Microsoft, USA

Keyword discovery is an unsupervised technology that can help to process collections of speech and capture repeated patterns. This technology becomes useful and provides solution for unsupervised content analysis tasks, especially when the acoustic and lexical characteristics are not known in advance or there is little or no data to model these characteristics via statistical models. In these situations, keyword discovery can find potentially important words for further analysis using minimal resources. Unfortunately, keyword discovery performance heavily depends on the quality of the features used to characterize the raw signal and the alignment algorithm used to find similar feature subsequences. It is not yet fully understood which features and alignment algorithms work well in different scenarios and for different tasks, and there are very few diagnostic techniques for improving our understanding. In this paper, we present two diagnostic measurements that can be used to directly assess the quality of alignments between sequences of features independently of the intended use of the alignments downstream. We argue that such diagnostic techniques are valuable for intrinsically assessing speech features and alignment algorithms for keyword detection.

Robust Retrieval Models for False Positive Errors in Spoken Documents
Sho Kawasaki, Tomoyosi Akiba; Toyohashi University of Technology, Japan

How to deal with speech recognition errors and out-of-vocabulary (OOV) words, which are referred to as false negative errors, are common challenges in spoken document processing. To deal with them in spoken content retrieval (SCR), the SCR method that incorporated spoken term detection (STD) as the pre-process stage (referred to as STD-SCR) has been proposed. However, the STD-SCR tends to increase false positive errors in compensation for reducing false negative errors. In this work, we propose robust retrieval models for false positive errors by using word co-occurrences. The words that co-occur in a given query are semantically related, so that they are likely to co-occur also in the document to be retrieved. On the other hand, if a word in a given query appears alone in a document, it is more like a false positive. We incorporate this idea into two retrieval models commonly used in the literature, i.e. the vector space model and the query likelihood model. Our experimental result showed our proposed extensions on the retrieval

NOTES
models successfully improved the retrieval performance not only for the STD-SCR but also for the conventional SCR method.

**Semantic Retrieval of Personal Photos Using Matrix Factorization and Two-Layer Random Walk Fusing Sparse Speech Annotations with Visual Features**

*Yuan-ming Liu, Yi-sheng Fu, Hung-yi Lee, Lin-shan Lee; National Taiwan University, Taiwan*

Wed-P-17-10, Poster

It is very attractive but technically very challenging if the user can retrieve photos from a huge collection using high-level personal queries (e.g., "Uncle Bill's House"). This paper proposes a set of approaches to achieve the goal assuming only 30% of the photos are annotated by sparse and spontaneously spoken descriptions when the photos are taken. We fuse the visual features (visual words plus global visual concepts from Columbia 374 detector) with the sparse speech features and train latent topics for the photos in the collection using non-negative matrix factorization. The retrieved results are then enhanced by two-layer mutually reinforced random walk based on different types of features. In this way it becomes possible to retrieve photos without speech annotation or with sparse annotations regarding different categories of information (e.g. where and who) because of the fused visual/speech features and jointly trained latent topics. Very encouraging results were obtained in initial experiments.

**Audio Thumbnails for Spoken Content Without Transcription Based on a Maximum Motif Coverage Criterion**

*Guillaume Gravier, Nathan Souviraì-Labastie, Sébastien Campion, Frédéric Bimbot; IRISA, France*

Wed-P-17-11, Poster

The paper presents a system to create audio thumbnails of spoken content, i.e., short audio summaries representative of the entire content, without resorting to a lexical representation. As an alternative to searching for relevant words and phrases in a transcript, unsupervised motif discovery is used to find short, word-like, repeating fragments at the signal level without acoustic models. The output of the word discovery algorithm is exploited via a maximum motif coverage criterion to generate a thumbnail in an extractive manner. A limited number of relevant segments are chosen within the data so as to include the maximum number of motifs while remaining short enough and intelligible. Evaluation is performed on broadcast news reports with a panel of human listeners judging the quality of the thumbnails. Results indicate that motif-based thumbnails stand between random thumbnails and ASR-based keywords, however still far behind thumbnails and keywords humanly authored.

**Semantically Based Search in a Social Speech Task**

*Fernando Garcia, Emilio Sanchís, Ferran Pla; Universidad Politécnica de Valencia, Spain*

Wed-P-17-12, Poster

In this work, we present an approach for semantically based search for similar segments of a speech task, i.e., the search for audio segments in a row audio repository that are semantically related to the audio segment given by a user. Our approach is based on the lexical representation of segments of words that are enriched by semantic relations. We have studied different distance measures and the lexical/semantic representation of the segments. We present experiments for a task of recorded dialogs between students talking about whatever they want, which is a semantically unbounded task. The results, which are encouraging, indicate the potential advantages of using this approach to address this problem.

**Study of Changes in Glottal Vibration Characteristics During Laughter**

*Vinay Kumar Mittal, B. Yegnanarayana; IIT Hyderabad, India*

Wed-P-18-1, Poster

Laughter in speech has been studied mostly relying upon the spectral representation like formants and harmonics derived from the short-time spectrum. Significant changes appear to take place in the characteristics of glottal source of excitation during the production of laughter, but these changes have not been explored much. In this study, we examine the changes in the glottal vibration characteristics in laughter production, using the electroglottograph (EGG) signals. The excitation source characteristics are also examined from the corresponding acoustic signal, using a modified zero-frequency filtering method. Changes are examined in three categories, namely, normal speech, laughed-speech and nonspeech-laugh. Results using both EGG and acoustic signals are similar. The closed phase quotient in each glottal cycle is observed to decrease more for nonspeech-laugh calls than laughed-speech, with reference to normal speech. Correspondingly, the instantaneous fundamental frequency, hence pitch, increases more for nonspeech-laugh.

**On Predicting the Unpleasantness Level of a Sound Event**

*Stavros Ntalampiras 1, Ilyas Potamitis 2; 1 European Commission, Italy; 2 TEI of Crete, Greece*

Wed-P-18-2, Poster

This work presents a novel framework for the automatic assessment of the unpleasantness caused by audio events to a human listener which is a relatively new research problem. Mel-frequency cepstral coefficients and temporal modulation parameters were employed to characterize 75 sound stimuli varying from animal calls to baby cries. The final assessment is made by means of a clustering scheme realized by Gaussian mixture models. The proposed framework leads to the best performance in terms of mean squared error and correlation between predicted and measured unpleasantness levels reported so far in the literature.

**Predicting When to Laugh with Structured Classification**

*Bilal Piot 1, Olivier Pietquin 2, Matthieu Geist 1; 1 Supélec, France; 2 LIFL, France*

Wed-P-18-3, Poster

Today, Embodied Conversational Agents (ECAs) are emerging as natural media to interact with machines. Applications are numerous and ECAs can reduce the technological gap between people by providing user-friendly interfaces. Yet, ECAs are still unable to produce social signals appropriately during their interaction with humans, which tends to make the interaction less instinctive.
Especially, very little attention has been paid to the use of laughter in human-avatar interactions despite the crucial role played by laughter in human-human interaction. In this paper, a method for predicting the most appropriate moment for laughing for an ECA is proposed. Imitation learning via a structured classification algorithm is used in this purpose and is shown to produce a behavior similar to humans’ on a practical application: the yes/no game.

Conversational Structures Affecting Auditory Likeability

Benjamin Weiss1, Katrin Schoenenberg2; 1T-Labs, Germany; 2Technische Universität Berlin, Germany

Wed-P-18-4, Poster

Three-person telephone conferences of unacquainted people are conducted by means of pre-defined scenarios providing individual information and goals. Interlocutors rate the likeability of each other after a training session as well as after the actual conference. The recordings of the conferences are manually annotated concerning speaker’s verbal contribution, pauses, and back-channels. Regression analysis reveals likeability ratings after the conference to be dominated by the ratings before the conversation, but simple parameters like number of turns also contribute significantly to a descriptive model. The regression model for ratings averaged for both interlocutors provide a similar fit as the one for pair-wise ratings. Exchanging of the manually obtained parameters by automatically estimated values still results in significant regressions, indicating facilitation for future research.

Towards the Adaptation of Prosodic Models for Expressive Text-to-Speech Synthesis

Mathieu Avanzi1, George Christodoulides2, Damien Lolive3, Elisabeth Delais-Roussarie1, Nelly Barbot1; 1LLF (UMR 7110), France; 2Université catholique de Louvain, Belgium; 3IRISA, France

Wed-P-18-5, Poster

This paper presents a preliminary study whose main aim is to characterize four distinct speaking styles according to a limited set of prosodic features, including the length of prosodic phrases (AP and IP), the distribution of stressed syllables, pitch register span, the duration of silent pauses, etc. The analysis was performed using semi-automatic procedures on a corpus consisting of 30 minutes of speech per style. The study focuses on four styles, all of which are "overtly addressed to a given audience", but differ as to the nature of the audience (adults vs. children) and the desired impact of the address ("importance of being understood and convincing, or not"). Data analysis reveals that (a) dictation (addressed to children) and political speeches (addressed to adults) are different to the two other speaking styles (reading of novels and fairy tales) with respect to a specific set of prosodic cues; while (b) the speeches addressed to children differ from the ones addressed to adults, with respect to another set of prosodic cues (especially pitch register span). These results have an interesting practical application: refining the design of pre-processing prosodic modules in a text-to-speech system, in order to improve the expressivity of synthesized speech.

Data-Driven Generation of Text Balloons Based on Linguistic and Acoustic Features of a Comics-Anime Corpus

Sho Matsumiya, Sakriani Sakti, Graham Neubig, Tomoki Toda, Satoshi Nakamura; NAIST, Japan

Wed-P-18-6, Poster

Most automatic speech recognition systems existing today are still limited to recognizing what is being said, without being concerned with how it is being said. On the other hand, research on emotion recognition from speech has recently gained considerable interest, but how those emotions could be expressed in text-based communication has not been widely investigated. Our long-term goal is to construct expressive speech-to-text systems that convey all information from acoustic speech, including verbal message, emotional state, speaker condition, and background noise, into unified text-based communication. In this preliminary study, we start with developing a system that can convey emotional speech into text-based communication by way of text balloons. As there exist many possible ways to generate the text balloons, we propose to utilize linguistic and acoustic features based on comic books and anime films. Experimental results reveal that expressive text is more preferable than static text, and the system is able to estimate the shape of text balloons with 87.01% accuracy.

Learning L2 Prosody is More Difficult Than You Realize — F0 Characteristics and Chunking Size of L1 English, TW L2 English and TW L1 Mandarin

Chiu-yu Tseng, Chao-yu Su; Academia Sinica, Taiwan

Wed-P-18-7, Poster

We compare the F0 output and chunking size of speech units of L1 English, TW L2 English and L1 Mandarin to see what some of the major intrinsic prosodic difference could be, what prosodic features could account for TW L2 English and in what way prosodic transfer occurs. Results show that the fundamental prosodic difference of the two languages is how in the pitch domain English requires sharper high/low contrast by higher-level prosodic units but less such contrast in accentuating lower-level prosodic units whereas Mandarin patterns are the exact opposite. Explanations are provided regarding how TW L2 English differs from L1, why prosodic transfer merits detailed analysis, and why mastering English prosody is especially difficult.

Investigating Prosodic Relations Between Initiating and Responding Laughs

Khiet P. Truong1, Jürgen Trouvain2; 1University of Twente, The Netherlands; 2Universität des Saarlandes, Germany

Wed-P-18-8, Poster

In dialogue, it is not uncommon for people to laugh together. This joint laughter often results in overlapping laughter, consisting of an initiating laugh (the first one), and a responding laugh (the second one). In previous studies, we found that overlapping laughs are acoustically different from non-overlapping ones. So far, we have considered overlapping laughs as one category. Consequently, it is unknown whether there are also acoustic differences between initiating laughs and responding laughs. In this paper, we make a distinction between initiating, responding, and non-overlapping laughs and compare their acoustic characteristics. In particular, we will investigate the prosodic relations between initiating and responding laughs. Do these relations point to a form of accommodation and mimicry? To what extent are initiating and responding laughs paired to each other? The analyses were performed on two
speech corpora containing spontaneous conversations between two speakers. Results show indications that initiating and responding laughs share several similar acoustic features that point toward accommodation and mimicry mechanisms.

**Application of Image Processing Methods to Filled Pauses Detection from Spontaneous Speech**

Dmytro Prylikpo, Olga Egorow, Ingo Siegert, Andreas Wendemuth; Otto-von-Guericke-Universität Magdeburg, Germany

To obtain a more human-like interaction with technical systems, those have to be adaptable to the users individual preferences, and current emotional state. In human-human interaction the behaviour of the speaker is characterised by semantic and prosodic cues, given (among other indicators) as short feedback signals. These so called filled pauses minimally convey certain dialogue functions such as attention, understanding, confirmation, or other attitudinal reactions. These signals play a valuable role in the progress and coordination of interaction. Hereby, the first step enabling an automatic system to react on these signals is the detection of them within the users utterances. This is a quite complex task, as the filled pauses are phonetically short, consisting mostly only of one vowel and one consonant. In this paper we present our methods to detect filled pauses in a naturalistic interaction utilising the LAST MINUTE corpus. We used an SVM classifier and improved the results further, by applying a Gaussian filter to infer temporal context information and performing a morphological opening to filter false alarms. We obtained recall of 70%, precision of 55%, and AUC of 0.94.

**Perception of Sentence Stress in English Infant Directed Speech**

Sofoklis Kakouros, Okko Räsänen; Aalto University, Finland

Various studies have examined the acoustic features in infant directed speech (IDS) and adult directed speech (ADS). However, there are few speech corpora with prominence annotation from multiple listeners or analysis of the acoustic properties of the stressed versus unstressed words, most studies and corpora focusing on syllabic stress. In order to fill this gap, the current study analyzes the acoustic properties of sentence stress in a corpus of English IDS. More specifically, the work is one of the first analyzing IDS as perceived by adult listeners, providing inter-annotator agreement ratings and an analysis of the acoustic correlates of sentence stress with regard to the most important prosodic features encountered in the literature: fundamental frequency, intensity, word duration, and spectral tilt. The analysis shows that all of the analyzed features correlate with the perception of stress, indicating that the sentential prominence in IDS is conveyed by similar acoustic characteristics that are known to be relevant for stress perception in ADS.

**Automatic Recognition of Attitudes in Video Blogs — Prosodic and Visual Feature Analysis**

Noor Alhusna Madzlan, JingGuang Han, Francesca Bonin, Nick Campbell; Trinity College Dublin, Ireland

This paper reports a study of attitude manifestations in video blogs. We describe the manual annotation of speaker attitudes in a corpus of over 130 video blogs and present an analysis of prosodic and visual cues in relation to attitude states. We use machine learning techniques for the automatic prediction of attitudes from prosodic and visual features in video blogs and compare the performance of prosodic and visual feature sets.

"Was That Your Mother on the Phone?": Classifying Interpersonal Relationships Between Dialog Participants with Lexical and Acoustic Properties

Denys Katerenchuk, David Guy Brizan, Andrew Rosenberg; CUNY Graduate Center, USA

Understanding interpersonal relationships provides important context in understanding spoken communication. In addition to increasing knowledge of the social indicators in spoken communication, the automatic recognition of interpersonal relationships has an application in providing structure to social networks. This paper presents exploratory work on the challenging problem of distinguishing family from friends in spontaneous dialogs drawn from the CALLHOME English corpus. We find both acoustic/prosodic and lexical features useful in classifying these relationships. In binary classification experiments, we achieve accuracy of 10.71% absolute improvement over chance (50%) assignment.

**Oral Session 25: Features and Robustness in Speaker and Language Recognition**

Garnet 213-218
13:30 - 15:30, Wednesday 17 September 2014
Chairs: Koichi Shinoda and Mitchell McLaren

**Combining Source and System Information for Limited Data Speaker Verification**

Rohan Kumar Das¹, Abhiram B.², S.R.M. Prasanna¹, A.G. Ramakrishnan²; IIT Guwahati, India; ²Indian Institute of Science, India

Speaker verification using limited data is always a challenge for practical implementation as an application. An analysis on speaker verification studies for an i-vector based method using Mel-Frequency Cepstral Coefficient (MFCC) feature shows that the performance drops drastically as the duration of test data is reduced. This decrease in performance is due to insufficient phonetic coverage when we capture only the vocal tract feature. However the same can be improved if some source characteristics are taken into consideration. This paper attempts to improve the speaker verification performance using source characteristics. A recently proposed characterization of the voice source signal called the discrete cosine transform of the integrated linear prediction residual (DCTILPR) has been found to be useful as a speaker-specific feature. Speaker verification is performed over short test utterances in the NIST 2003 database using both the DCTILPR and MFCC features, and their score-level combination is found to give a significant improvement. The best performance is achieved when using only the MFCC features.

**New Insight into the Use of Phone Log-Likelihood Ratios as Features for Language Recognition**

Mireia Diez, Amparo Varona, Mikel Penaquirkano, Luis Javier Rodriguez-Fuentes, German Bordel; Universidad del Pais Vasco, Spain

Phone Log-Likelihood Ratio (PLLR) features have been recently introduced as an effective way of making use of frame-level phone
posterior probabilities in language and speaker recognition systems. In this paper, a deep insight into PLLR features is made and further evidence of the usefulness of these features in spoken language recognition tasks is provided, with a new set of experiments carried out on the NIST 2007 LRE dataset, combining the latest progresses made in optimizing the features. PLLR features are projected into a subspace that enhances the information retrieved by the system. Then, dimensionality reduction is performed on the projected subspace by means of Principal Component Analysis, and shifted deltas are computed on the reduced features to optimize performance. Figures attained are among the best reported so far on the NIST 2007 LRE dataset.

Robust Language Identification Using Convolutional Neural Network Features

Sriram Ganapathy\textsuperscript{1}, Kyu Han\textsuperscript{1}, Samuel Thomas\textsuperscript{1}, Mohamed Omar\textsuperscript{1}, Maarten Van Segbroeck\textsuperscript{2}, Shrikanth S. Narayanan\textsuperscript{2}; \textsuperscript{1}IBM T.J. Watson Research Center, USA; \textsuperscript{2}University of Southern California, USA

The language identification (LID) task in the Robust Automatic Transcription of Speech (RATS) program is challenging due to the noisy nature of the audio data collected over highly degraded radio communication channels as well as the use of short duration speech segments for testing. In this paper, we report the recent advances made in the RATS LID task by using bottleneck features from a convolutional neural network (CNN). The CNN, which is trained with labelled data from one of target languages, generates bottleneck features which are used in a Gaussian mixture model (GMM)-i-vector LID system. The CNN bottleneck features provide substantial complimentary information to the conventional acoustic features even on languages not seen in its training. Using these bottleneck features in conjunction with acoustic features, we obtain significant improvements (average relative improvements of 25\% in terms of equal error rate (EER) compared to the corresponding acoustic system) for the LID task. Furthermore, these improvements are consistent for various choices of acoustic features as well as speech segment durations.

Acoustic Feature Transformation Using UBM-Based LDA for Speaker Recognition

Chengzhu Yu, Gang Liu, John H.L. Hansen; University of Texas at Dallas, USA

In state-of-the-art speaker recognition system, universal background model (UBM) plays a role of acoustic space division. Each Gaussian mixture of trained UBM represents one distinct acoustic region. The posterior probabilities of features belonging to each region are further used as core components of Baum-Welch statistics. Therefore, the quality of estimated Baum-Welch statistics depends highly on how acoustic regions are separable with each other. In this paper, we propose to transform the front end acoustical features into a space where the separability of mixtures of trained UBM can be optimized. To achieve this, an UBM was first trained from the acoustical features and a transformation matrix is estimated using linear discriminant analysis (LDA) by treating each mixture of trained UBM as independent class. Therefore, the proposed method named as UBM-based LDA (uLDA) does not require any speaker labels or other supervised information. The obtained transformation matrix is then applied to acoustic features for i-Vector extraction. Experimental results on the male part of core conditions of NIST SRE 2010 dataset confirmed the improved performance using proposed method.

SNR-Dependent Mixture of PLDA for Noise Robust Speaker Verification

Man-Wai Mak; Hong Kong Polytechnic University, China

This paper proposes a mixture of SNR-dependent PLDA models to provide a wider coverage on the i-vector spaces so that the resulting i-vector/PLDA system can handle test utterances with a wide range of SNR. To maximise the coordination among the PLDA models, they are trained simultaneously via an EM algorithm using utterances contaminated with noise at various levels. The contribution of a training i-vector to individual PLDA models is determined by the posterior probability of the utterance’s SNR. Given a test i-vector, the marginal likelihoods from individual PLDA models are linear combined based on the the posterior probabilities of the test utterance and the targetspeaker’s utterance. Verification scores are the ratio of the marginal likelihoods. Results based on NIST 2012 SRE suggest that this soft-decision scheme is particularly suitable for the situations where the test utterances exhibit a wide range of SNR.

Nearest Neighbor Discriminant Analysis for Robust Speaker Recognition

Seyed Omid Sadjadi, Jason Pelecanos, Weizhong Zhu; \textit{IBM T.J. Watson Research Center, USA}

With the advent of i-vectors, linear discriminant analysis (LDA) has become an integral part of many state-of-the-art speaker recognition systems. Here, LDA is primarily employed to annihilate the non-speaker related (e.g., channel) directions, thereby maximizing the inter-speaker separation. The traditional approach for computing the LDA transform uses parametric representations for both intra- and inter-speaker scatter matrices that are based on the Gaussian distribution assumption. However, it is known that the actual distribution of i-vectors may not necessarily be Gaussian, and in particular, in the presence of noise and channel distortions. Motivated by this observation, we present an alternative non-parametric discriminant analysis (NDA) technique that measures both the within- and between-speaker variation on a local basis using the nearest neighbor rule. The effectiveness of the NDA method is evaluated in the context of noisy speaker recognition tasks using speech material from the DARPA Robust Automatic Transcription of Speech (RATS) program. Experimental results indicate that the NDA is more effective than the traditional parametric LDA for speaker recognition under noisy and channel degraded conditions.

Notes
Incorporating Lexical and Prosodic Information at Different Levels for Meeting Summarization

Catherine Lai, Steve Renals; University of Edinburgh, UK

Wed O-26-3

This paper investigates how prosodic features can be used to augment lexical features for meeting summarization. Automatic detection of summary-worthy content using non-lexical features, like prosody, has generally focused on features calculated over dialogue acts. However, a salient role of prosody is to distinguish important words within utterances. To examine whether including more fine grained prosodic information can help extractive summarization, we perform experiments incorporating lexical and prosodic features at different levels. For ICSI and AMI meeting corpora, we find that combining prosodic and lexical features at a lower level has better AUROC performance than adding in prosodic features derived over dialogue acts. ROUGE F-scores also show the same pattern for the ICSI data. However, the differences are less clear for the AMI data where the range of scores is much more compressed. In order to understand the relationship between the generated summaries and differences in standard measures, we look at the distribution of extracted content over meeting as well as summary redundancy. We find that summaries based on dialogue act level prosody better reflect the amount of human annotated summary content in meeting segments, while summaries derived from prosodically augmented lexical features exhibit less redundancy.

Subspace Gaussian Mixture Models for Dialogues Classification

Mohamed Bouallegue, Mohamed Morchid, Richard Dufour, Driss Matrouf, Georges Linarès, Renato De Mori; LIA, France

Wed O-26-4

The main objective of this paper is to identify themes from dialogues of telephone conversations in a real-life customer care service. In order to capture significant semantic content in spite of high expression variability, features are extracted in a large number of hidden spaces constructed with a Latent Dirichlet Allocation (LDA) approach. Multiple views of a spoken document can then be represented with several hidden topic models. Nonetheless, the model diversity due to the multi-model approach introduces a new type of variability. An approach is proposed based on features extracted in a common homogenous subspace with the purpose of reducing the multi-span representation variability. A Gaussian Mixture Model subspace model, inspired by previous work on speaker identification, is proposed for theme identification. This representation, novel for theme classification, is compared with the direct application of multiple topic-model representations. Experiments are reported using a corpus collected in the call center of the Paris Transportation Service. Results show the effectiveness of the proposed representation paradigm with a theme identification accuracy of 78.8%, showing a significant improvement with respect to previous results on the same corpus.

Factor Analysis Based Semantic Variability Compensation for Automatic Conversation Representation

Mohamed Bouallegue, Mohamed Morchid, Richard Dufour, Driss Matrouf, Georges Linarès, Renato De Mori; LIA, France

Wed O-26-5

The main objective of this paper is to identify themes from di-

NOTES
alogues of telephone conversations in a real-life customer care service. In this task, the word semantic variability contained in these conversations may impact the classification performance by retaining the noise in their vectorial representation. In this article, we propose an original method to compensate this semantic variability using the Factor Analysis (FA) paradigm, initially designed for speech processing tasks to compensate the acoustic variability, mainly in Speaker Verification (SV) and Automatic Speech Recognition (ASR). In our proposal, we used the FA paradigm to estimate the semantic variability as an additive component located in a subspace of low dimension (with respect to the super-vector space). This additive semantic variability is estimated in Factor Analysis model space. From this estimation, a specific vector transformation is obtained and is applied to vectors of dialogue representation. Experiments are reported using a corpus collected in the call center of the Paris Transportation Service. Results show the effectiveness of the proposed representation paradigm with a theme identification accuracy of 80.0%, showing a significant improvement with respect to previous results on the same corpus.

Speech Cohesion for Topic Segmentation of Spoken Contents
Abdessalam Bouchekif, Géraldine Damnati, Delphine Charlet; Orange Labs, France

In this paper, we introduce the notion of speech cohesion for topic segmentation of a spoken content. The aim is to integrate speaker information and lexical information within a single cohesion value. Based on a lexical cohesion system, we propose an approach that directly integrates the speaker distribution when processing the cohesion. A potential boundary is effective if the joint distribution of terms and speakers is different enough from one side of the boundary to the other. Beyond speaker distribution, we also propose to take into account speaker identification and to confront speaker identities to identities mentioned in the spoken content in order to reinforce cohesion of a topic segment. Experiments run on three corpora of various Broadcasts News formats collected from 9 French TV channels, show a significant improvement in the overall topic segmentation process.

Oral Session 27: DNN Learning
Peridot 204-205
13:30 – 15:30, Wednesday 17 September 2014
Chairs: Frank Seide and Khe Chai Sim

A Comparative Analytic Study on the Gaussian Mixture and Context Dependent Deep Neural Network Hidden Markov Models
Yan Huang, Dong Yu, Chaojun Liu, Yifan Gong; Microsoft, USA

We conducted a comparative analytic study on the context-dependent Gaussian mixture hidden Markov model (CD-GMM-HMM) and deep neural network hidden Markov model (CD-DNN-HMM) with respect to the phone discrimination and the robustness performance. We found that the DNN can significantly improve the phone recognition performance for every phoneme with 15.6% to 39.8% relative phone error rate reduction (P ERR). It is particularly good at discriminating certain consonants, which are found to be “hard” in the GMM. On the robustness side, the DNN outperforms the GMM at all SNR levels, across different devices, and under all speaking rate with nearly uniform improvement. The performance gap with respect to different SNR levels, distinct channels, and varied speaking rate remains large. For example, in CD-DNN-HMM, we observed 1 – 2% performance degradation per 1dB SNR drop; 20 – 25% performance gap lead the best and least well performed devices; 15 – 30% relative word error rate increase when the speaking rate speeds up or slows down by 30% from the “sweet” spot. Therefore, we conclude the robustness remains to be a major challenge in the deep learning acoustic model. Speech enhancement, channel normalization, and speaking rate compensation are important research areas in order to further improve the DNN model accuracy.

Asynchronous, Online, GMM-Free Training of a Context Dependent Acoustic Model for Speech Recognition
Michiel Bacchiani, Andrew Senior, Georg Heigold; Google, USA

We propose an algorithm that allows online training of a context dependent DNN model. It designs a state inventory based on DNN features and jointly optimizes the DNN parameters and alignment of the training data. The process allows flat starting a model from scratch and avoids any dependency on a GMM acoustic model to bootstrap the training process. A 15k state model trained with the proposed algorithm reduced the error rate on a mobile speech task by 24% compared to a system bootstrapped from a CI GMM and by 16% compared to a system bootstrapped from a CD GMM system.

Autoregressive Product of Multi-Frame Predictions Can Improve the Accuracy of Hybrid Models
Navdeep Jaitly1, Vincent Vanhoucke2, Geoffrey Hinton1, 2University of Toronto, Canada; 2Google, USA

We describe a simple but effective way of using multi-frame targets to improve the accuracy of Artificial Neural Network-Hidden Markov Model (ANN-HMM) hybrid systems. In this approach a Deep Neural Network (DNN) is trained to predict the forced-alignment state of multiple frames using a separate softmax for each of the frames. This is in contrast to the usual method of training a DNN to predict only the state of the central frame. By itself this is not sufficient to improve accuracy of the system significantly. However, if we average the predictions for each frame — from the different contexts it is associated with — we achieve state of the art results on TIMIT using a fully connected Deep Neural Network without convolutional architectures or dropout training. On a 14 hour subset of Wall Street Journal (WSJ) using a context dependent DNN-HMM system it leads to a relative improvement of 6.4% on the dev set (test-dev93) and 9.3% on test set (test-eval92).

Learning Small-Size DNN with Output-Distribution-Based Criteria
Jinyu Li1, Rui Zhao2, Jui-Ting Huang1, Yifan Gong1; 1Microsoft, USA; 2Microsoft, China

Deep neural network (DNN) obtains significant accuracy improvements on many speech recognition tasks and its power comes from the deep and wide network structure with a very large number of parameters. It becomes challenging when we deploy DNN on devices which have limited computational and storage resources. The common practice is to train a DNN with a small number of hidden nodes and a small senone set using the standard training process, leading to significant accuracy loss. In this study, we
propose to better address these issues by utilizing the DNN output distribution. To learn a DNN with small number of hidden nodes, we minimize the Kullback-Leibler divergence between the output distributions of the small-size DNN and a standard large-size DNN by utilizing a large number of un-transcribed data. For better senone set generation, we cluster the senones in the large set into a small one by directly relating the clustering process to DNN parameters, as opposed to decoupling the senone generation and DNN training process in the standard training. Evaluated on a short message dictation task, the proposed two methods get 5.08% and 1.33% relative word error rate reduction from the standard training method, respectively.

**Ensemble Deep Learning for Speech Recognition**

*Li Deng, John C. Platt; Microsoft, USA*

Deep learning systems have dramatically improved the accuracy of speech recognition, and various deep architectures and learning methods have been developed with distinct strengths and weaknesses in recent years. How can ensemble learning be applied to these varying deep learning systems to achieve greater recognition accuracy is the focus of this paper. We develop and report linear and log-linear stacking methods for ensemble learning with applications specifically to speech-class posterior probabilities as computed by the convolutional, recurrent, and fully-connected deep neural networks. Convex optimization problems are formulated and solved, with analytical formulas derived for training the ensemble-learning parameters. Experimental results demonstrate a significant increase in phone recognition accuracy after stacking the deep learning subsystems that use different mechanisms for computing high-level, hierarchical features from the raw acoustic signals in speech.

**Learning Conditional Random Field with Hierarchical Representations for Dialogue Act Recognition**

*Yucan Zhou¹, Qinghua Hu¹, Jie Liu², Yuan Jia³;¹ Tianjin University, China;² Nankai University, China;³ Chinese Academy of Social Sciences, China*

The analysis of dialogue act is important for computers to understand natural-language dialogues because the dialogue act of an utterance characterizes the speaker's intention. In this paper, we create a new model that adapts Conditional Random Field (CRF) with efficient hierarchical representations of the raw inputs to solve the dialogue act recognition problem. The proposed model has two advantages. First, CRF can model the statistical dependencies between the utterances which carry important information to determine the dialogue act label of an utterance. Second, the hierarchical representations potentially contain some more abstract concepts with greater predictive power. To verify the effectiveness of our model, we compare it with several baseline methods on a dialogue act classification task. The results of the experiments demonstrate that our model performs much better than all the baseline methods.

---

**Oral Session 28: Perception of Emotion and Prosody**

**Peridot 201**

13:30 - 15:30, Wednesday 17 September 2014

**Chairs:** Takao Kobayashi and Jens Edlund

---

**Can Adolescents with Autism Perceive Emotional Prosody?**

*Cristiane Hsu, Yi Xu; University College London, UK*

*Wed O-28-1*

Past findings on the perception of emotional prosody by individuals with Autism Spectrum Disorder (ASD) have been incongruent, and a main reason is the lack of clarity about the contributions of specific acoustic features to emotion perception. In this study we test the perception of emotional prosody by adolescents with ASD using a recently developed prosody control method based on a bio-informational dimensions (BID) theory of emotion expressions. We synthesized a Mandarin sentence with different voice qualities using an articulatory synthesizer, and then acoustically manipulated their formant dispersion, median pitch and pitch range. With these utterances we compared the ability to perceive body size, emotion and attitude by high-functioning adolescents with ASD, typically developing adolescents and young adults. Results showed that the three groups made similar perceptual judgements, but the sensitivity of adolescents with ASD to the acoustic manipulations was lower than their typically developing peers, who in turn exhibited less sensitivity than young adults. These findings show that individuals with ASD have a reduced rather than a total lack of ability to perceive emotional prosody, suggesting a delay in their developmental trajectory. The findings also demonstrate the effectiveness of the BID-based method in testing perception of emotional prosody.

**Age, Hearing Loss and the Perception of Affective Utterances in Conversational Speech**

*Juliane Schmidt, Esther Janse, Odette Scharenborg; Radboud Universiteit Nijmegen, The Netherlands*

*Wed O-28-2*

This study investigates whether age and/or hearing loss influence the perception of the emotion dimensions arousal (calm vs. aroused) and valence (positive vs. negative attitude) in conversational speech fragments. Specifically, this study focuses on the relationship between participants’ ratings of affective speech and acoustic parameters known to be associated with arousal and valence (mean F0, intensity, and articulation rate). Ten normally-hearing younger and ten older adults with varying hearing loss were tested on two rating tasks. Stimuli consisted of short sentences taken from a corpus of conversational affective speech. In both rating tasks, participants estimated the value of the emotion dimension at hand using a 5-point scale. For arousal, higher intensity was generally associated with higher arousal in both age groups. Compared to younger participants, older participants rated the utterances as less aroused, and showed a smaller effect of intensity on their arousal ratings. For valence, higher mean F0 was associated with more negative ratings in both age groups. Generally, age group differences in rating affective utterances may not relate to age group differences in hearing loss, but rather to other differences between the age groups, as older participants’ rating patterns were not associated with their individual hearing loss.
In interpersonal interactions, speech and body gesture channels are internally coordinated towards conveying communicative intentions. The speech-gesture relationship is influenced by the internal emotion state underlying the communication. In this paper, we focus on uncovering the emotional effect on the interrelation between speech and body gestures. We investigate acoustic features describing speech prosody (pitch and energy) and vocal tract configuration (MFCCs), as well as three types of body gestures, viz., head motion, lower and upper body motions. We employ mutual information to measure the coordination between the communicative channels, and analyze the quantified speech-gesture link with respect to distinct levels of emotion attributes, i.e., activation and valence. The results reveal that the speech-gesture coupling is generally tighter for low-level activation and high-level valence, compared to high-level activation and low-level valence. We further propose a framework for modeling the dynamics of speech-gesture interaction. Experimental studies suggest that such quantified coupling representations can well discriminate different levels of activation and valence, reinforcing that emotions are encoded in the dynamics of the multimodal link. We also verify that the structures of the coupling representations are emotion-dependent using subspace-based analysis.

When Voices Get Emotional: A Study of Emotion-Enhanced Memory and Impairment During Emotional Prosody Exposure

Cyrielle Chappuis, Didier Grandjean; Université de Genève, Switzerland

Perception of Pitch Tails at Potential Turn Boundaries in Swedish

Margaret Zellers; KTH, Sweden

Towards a Perceptual Model of Speech Rhythm: Integrating the Influence of f0 on Perceived Duration

Robert Fuchs; Westfälische Wilhelms-Universität Münster, Germany

Notes
such as the F0 and voiced/unvoiced (V/UV) flag, resulting in a vibrating pitch contour in the synthesized voice. To address this problem, this paper investigates the use of an extra weighting vector on the acoustic output layer of the MD-DBN. It reduces the dimensional imbalance between spectrum and pitch parameters by giving different weighting coefficients to the spectrum, F0 and the V/UV flag in the training procedure. Experimental results show that wMD-DBN can generate smoother pitch contours and improve the naturalness of the synthesized speech.

TTS Synthesis with Bidirectional LSTM Based Recurrent Neural Networks

Yuchen Fan¹, Yao Qian², Feng-Long Xie², Frank K. Soong³; ¹Shanghai Jiao Tong University, China; ²Microsoft, China

Feed-forward, Deep neural networks (DNN)-based text-to-speech (TTS) systems have been recently shown to outperform decision-tree clustered context-dependent HMM TTS systems [1, 4]. However, the long time span contextual effect in a speech utterance is still not easy to accommodate, due to the intrinsic, feed-forward nature in DNN-based modeling. Also, to synthesize a smooth speech trajectory, the dynamic features are commonly used to constrain speech parameter trajectory generation in HMM-based TTS [2]. In this paper, Recurrent Neural Networks (RNNs) with Bidirectional Long Short Term Memory (BLSTM) cells are adopted to capture the correlation or co-occurrence information among two instants in a speech utterance for parametric TTS synthesis. Experimental results show that a hybrid system of DNN and BLSTM-RNN, i.e., lower hidden layers with a feed-forward structure which is cascaded with upper hidden layers with a bidirectional RNN structure of LSTM, can outperform either the conventional, decision tree-based HMM, or a DNN TTS system, both objectively and subjectively. The speech trajectory generated by the BLSTM-RNN TTS is fairly smooth and no dynamic constraints are needed.

Deep Neural Network Based Trainable Voice Source Model for Synthesis of Speech with Varying Vocal Effort

Tuomo Raitio¹, Antti Sunti¹, Lauri Juvela¹, Martti Vainio², Paavo Alku¹; ¹Aalto University, Finland; ²University of Helsinki, Finland

This paper studies a deep neural network (DNN) based voice source modelling method in the synthesis of speech with varying vocal effort. The new trainable voice source model learns a mapping between the acoustic features and the time-domain pitch-synchronous glottal flow waveform using a DNN. The voice source model is trained with various speech material from breathed, normal, and Lombard speech. In synthesis, a normal voice is first adapted to a desired style, and using the flexible DNN-based voice source model, a style-specific excitation waveform is automatically generated based on the adapted acoustic features. The proposed voice source model is compared to a robust and high-quality excitation modelling method based on manually selected mean glottal flow pulses for each vocal effort level and using a spectral matching filter to correctly match the voice source spectrum to a desired style. Subjective evaluations show that the proposed DNN-based method is rated comparable to the baseline method, but avoids the manual selection of the pulses and is computationally faster than a system using a spectral matching filter.

An Introduction to Computational Networks and the Computational Network Toolkit (Invited Talk)

Dong Yu, Adam Eversole, Michael L. Seltzer, Kaisheng Yao, Brian Guenter, Oleksii Kuchaiev, Frank Seide, Huaming Wang, Jasha Drappo, Zhiheng Huang, Geoff Zweig, Chris Rossbach, Jon Currey

We introduce the computational network (CN), a generalization of popular machine learning models such as deep neural network, recurrent neural network, convolutional neural network, and log linear model that can be expressed as a series of computation steps. We describe the benefits of such generalization and the key operations on the CN.

We further introduce the computational network toolkit (CNTK), a general purpose C++ implementation of computational networks. We describe its architecture and core functionalities and demonstrate that it can construct and learn models of arbitrary topology, connectivity, and recurrence. The toolkit will be released under a modified Microsoft Research license agreement for non-commercial use.

Poster Session 19: Speech Analysis and Perception

Max Atria Gallery
13:30 - 15:30, Wednesday 17 September 2014
Chair: Carlos Ishi

Acoustic Investigation of /th/ Lenition in Brunei Mandarin

Shufang Xu; Universiti Brunei Darussalam, Brunei

Wed-P-19-1, Poster

This study investigates the acoustic characteristics of /th/ lenition in conversational speech of Brunei Mandarin, a variety of Mandarin Chinese. Based on data from 20 Chinese Bruneians, /th/ lenition was found in the third-person pronoun tā /tʰa/, which is frequently pronounced as há [ha]. Perceptual judgments, spectrographic analysis and acoustic measurements were conducted to examine the features of this sound change. In comparison with the perceptual judgments, it was found that the spectrographic inspection yielded 83.6% correct classification of [tʰ] and [h] for female speakers and 77.2% for male speakers, indicating there is reasonably high reliability in identification in terms of spectral properties. Results of the acoustic measurements showed that there is an increase in high frequency intensity after the release of the closure for [h]. The results showed that the lack of burst and little increase in intensity are reasonably reliable cues for stop lenition.

Mapping Emotions into Acoustic Space: The Role of Voice Quality

Ting Wang¹, Hongwei Ding¹, Jianjing Kuang², Qiww Ma¹; ¹Tongji University, China; ²University of Pennsylvania, USA

Wed-P-19-2, Poster

This study aims to evaluate the importance of phonation cues in the acoustic realization of four vocal emotions (happiness, fear, anger, and sadness) in Mandarin Chinese. Our perception experiment confirmed that native listeners could differentiate these emotions. To explain how listeners accomplished the task in the perception experiment, we investigated the acoustic cues

NOTES
used by speakers. An acoustic analysis revealed that prosodic measures such as F₀, intensity, and duration failed to separate certain emotions in the acoustic space, and thus could not explain how native listeners can perceive the different emotions. However, by incorporating phonation-related cues in the multi-dimensional scaling acoustic space, different emotions were separated clearly. Principal components analysis further revealed the specific contribution of each acoustic measure. These results also allow some preliminary conjectures on how these acoustic components might represent the underlying emotion dimensions, namely, arousal, valence, and control.

**Principal Components of Auditory Spectro-Temporal Receptive Fields**

Nagaraj Mahajan¹, Nima Mesgarani², Hynek Hermansky¹; ¹Johns Hopkins University, USA; ²Columbia University, USA

*Wed-P-19-3, Poster*

More than two thousand auditory cortical spectro-temporal receptive fields (STRFs) of the ferret were analysed by Principal Component Analysis (PCA) to reveal their dominant properties. Results show that cortical levels of mammalian auditory processing enhance relatively low modulation spectral components of the signal around 3 Hz, using relatively broad spectral processing channels of the order of octave or more, evaluating mean spectral values and local spectral slopes within such frequency channels. These observations are consistent with some engineering techniques shown to be effective in machine recognition of speech.

**Segmentation in Singer Turns with the Bayesian Information Criterion**

Marwa Thlithi, Thomas Pellegrini, Julien Pinquier, Régine André-Obrecht; IRIT, France

*Wed-P-19-4, Poster*

As part of a project on indexing ethno-musicological audio recordings, segmentation in singer turns automatically appeared to be essential. In this article, we present the problem of segmentation in singer turns of musical recordings and our first experiments in this direction by exploring a method based on the Bayesian Information Criterion (BIC), which are used in numerous works in audio segmentation, to detect singer turns. The BIC penalty coefficient was shown to vary when determining its value to achieve the best performance for each recording. In order to avoid the decision about which single value is best for all the documents, we propose to combine several segmentations obtained with different values of this parameter. This method consists of taking a posteriori decisions on which segment boundaries are to be kept. A gain of 7.1% in terms of F-measure was obtained compared to a standard decision on which segment boundaries are to be kept. A gain of 7.1% in terms of F-measure was obtained compared to a standard decision on which segment boundaries are to be kept. Across the combined speaker set the first two PCs accounted for 60.5% of the variance, and remained related to phonetic height and backness. This can clearly be seen when the transformed area function data is compared to the first two vocal tract resonances and speech formants.

A Next Step Towards Measuring Perceived Quality of Speech Through Physiology

Sebastian Arndt¹, Markus Wenzel², Jan-Niklas Antons¹, Friedemann Köster¹, Sebastian Möller¹, Gabriel Curio¹; ¹T-Labs, Germany; ²Technische Universität Berlin, Germany; ³Charite-Universitätsmedizin, Germany

*Wed-P-19-6, Poster*

For service providers, it is important to know the perceived quality of a transmitted speech signal, as this can be an indicator for choosing one operator over the other. To evaluate the perceived quality, several methods have been introduced in the area of Quality of Experience (QoE), such as subjective, instrumental and physiological methods. In the current study, we expose the test participants towards several speech files, while brain activity was recorded with electroencephalography (EEG). Additionally, a subjective quality judgment is obtained after each presentation which is summarized as the mean opinion score (MOS). To validate the selected stimulus corpus, a quality prediction algorithm is used which calculates quality scores based on the speech signal. The recorded EEG data were set in relation to the subjective data and show promising results. In contrast to most previous studies in the domain of QoE using EEG, the study at hand uses a standardized test paradigm proposed by the International Telecommunication Unit (ITU).

Effect of Spectral Degradation to the Intelligibility of Vowel Sentences

Fei Chen, Sharon W.K. Wong, Lena L.N. Wong; University of Hong Kong, China

*Wed-P-19-7, Poster*

Based on the noise-replacement paradigm, recent studies showed that vowels carried more perceptual information for sentence intelligibility than consonants. Considering that vowels contain many important acoustic cues for speech perception, this study further assessed the effect of spectral degradation to the intelligibility of Mandarin vowel sentences. Mandarin sentences were processed to generate three types of spectrally degraded [i.e., fundamental frequency (F₀) flattened, sinewave synthesized, and noise-voiced] stimuli. Noise-replacement paradigm was implemented to preserve different amounts of vowel centers and replace the rest with noise. Listening experiments showed that flattening F₀ had a minimal effect on the intelligibility of Mandarin vowel sentences, and the harmonic structure within vowels accounted more for the intelligibility of Mandarin vowel sentences. While deleting vowel edges had little influence on the intelligibility of the unprocessed vowel sentences, it had a significantly negative effect on the intelligibility of vowel sentences with spectral degradation.

### Notes

- For service providers, it is important to know the perceived quality of a transmitted speech signal, as this can be an indicator for choosing one operator over the other. To evaluate the perceived quality, several methods have been introduced in the area of Quality of Experience (QoE), such as subjective, instrumental and physiological methods. In the current study, we expose the test participants towards several speech files, while brain activity was recorded with electroencephalography (EEG). Additionally, a subjective quality judgment is obtained after each presentation which is summarized as the mean opinion score (MOS). To validate the selected stimulus corpus, a quality prediction algorithm is used which calculates quality scores based on the speech signal. The recorded EEG data were set in relation to the subjective data and show promising results. In contrast to most previous studies in the domain of QoE using EEG, the study at hand uses a standardized test paradigm proposed by the International Telecommunication Unit (ITU).

- Based on the noise-replacement paradigm, recent studies showed that vowels carried more perceptual information for sentence intelligibility than consonants. Considering that vowels contain many important acoustic cues for speech perception, this study further assessed the effect of spectral degradation to the intelligibility of Mandarin vowel sentences. Mandarin sentences were processed to generate three types of spectrally degraded [i.e., fundamental frequency (F₀) flattened, sinewave synthesized, and noise-voiced] stimuli. Noise-replacement paradigm was implemented to preserve different amounts of vowel centers and replace the rest with noise. Listening experiments showed that flattening F₀ had a minimal effect on the intelligibility of Mandarin vowel sentences, and the harmonic structure within vowels accounted more for the intelligibility of Mandarin vowel sentences. While deleting vowel edges had little influence on the intelligibility of the unprocessed vowel sentences, it had a significantly negative effect on the intelligibility of vowel sentences with spectral degradation.
Consonant Context Effects on Vowel Sensorimotor Adaptation

Jeff Berry, John Jaeger IV, Melissa Wiedenhoef, Brittany Bernal, Michael T. Johnson; Marquette University, USA

Wed-P-19-8, Poster

Speech sensorimotor adaptation is the short-term learning of modified articulator movements evoked through sensory-feedback perturbations. A common experimental method manipulates acoustic parameters, such as formant frequencies, using real-time resynthesis of the participant’s speech to perturb auditory feedback. While some studies have examined phrases comprised of vowels, diphthongs, and semivowels, the bulk of research on auditory feedback-driven sensorimotor adaptation has focused on vowels in neutral contexts (/hVd/). The current study investigates coarticulatory influences of adjacent consonants on sensorimotor adaptation. The purpose is to evaluate differences in the adaptation effects for vowels in consonant environments that vary by place and manner of articulation. In particular, we addressed the hypothesis that contexts with greater intra-articulator coarticulation and more static articulatory postures (alveolars and fricatives) offer greater resistance to vowel adaptation than contexts with primarily inter-articulator coarticulation and more dynamic articulatory patterns (bilabials and stops). Participants completed formant perturbation-driven vowel adaptation experiments for varying CVCs. Results from discrete formant measures at the vowel midpoint were generally consistent with the hypothesis. Analyses of more complete formant trajectories suggest that adaptation can also (or alternatively) influence formant onsets, offsets, and transitions, resulting in complex formant pattern changes that may reflect modifications to consonant articulation.

Assessing Objective Characterizations of Phonetic Convergence

Gérard Bailly, Amélie Martin; GIPSA, France

Wed-P-19-9, Poster

This paper focuses on the study of the convergence between characteristics of speech segments — i.e., spectral characteristics of speech sounds — during live interactions between speaking dyads. The interaction data has been collected using an original verbal game called ‘verbal dominoes’ that provides a dense sampling of the acoustic spaces of the interlocutors. Two methods for characterizing phonetic convergence are here compared. The first one is based on a fine-grained analysis of the spectra of central frames of vowels (LDA) while the second one uses a more global speaker recognition technique (LLR). We show that convergence rates calculated by the two techniques correlate as the number of dominoes increases and that the LDA method well resists to the decrease of training and test material. We finally comment the impact of several factors on the computed convergence rates, i.e. interlocutors’ familiarity and sex pairs.

Generalizing Time-Frequency Importance Functions Across Noises, Talkers, and Phonemes

Michael I Mandel, Sarah E Yoho, Eric W Healy; Ohio State University, USA

Wed-P-19-10, Poster

Listeners can reliably identify speech in noisy conditions, although it is generally not known what specific features of speech are used to do this. We utilize a recently introduced data-driven framework to identify these features. By analyzing listening-test results involving the same speech utterance mixed with many different noise instances, the framework is able to compute the importance of each time-frequency point in the utterance to its intelligibility. This paper shows that a trained model resulting from this framework can generalize to new conditions, successfully predicting the intelligibility of novel mixtures. First, it can generalize to novel noise instances after being trained on mixtures involving the same speech utterance but different noises. Second, it can generalize to novel talkers after being trained on mixtures involving the same syllables produced by different talkers in different noises. Finally, it can generalize to novel phonemes, after being trained on mixtures involving different consonants produced by the same or different talkers in different noises. Aligning the clean utterances in time and then propagating this alignment to the features used in the intelligibility prediction improves this generalization performance further.

Does Elderly Speech Recognition in Noise Benefit from Spectral and Visual Cues?

Yatin Mahajan, Jeesun Kim, Chris Davis; University of Western Sydney, Australia

Wed-P-19-11, Poster

Previous research with young adults has shown that temporal (amplitude modulated, AM) cues are sufficient for recognizing speech in quiet but not for speech in noise. Speech perception in noise is more robust when spectral (frequency modulated, FM) cues are provided in addition to AM ones; visual cues (AV) provide an additional benefit. The elderly typically have problems recognizing speech in noise and it has recently been found that FM discrimination is worse in this group. Given this, it may be that elderly participants will not show an FM speech in noise benefit. To test the relative effectiveness of adding FM cues to AM ones for elderly versus young participants we compared auditory only (AO) speech identification of sentences, vowels and consonants in noise with AM and AM+FM presentation conditions. We also evaluated the relative effectiveness of visual cues for the elderly group. Although the elderly had poorer speech recognition performance overall, they showed a comparable visual benefit to the young group. Moreover, contrary to the prediction of a reduced benefit for FM cues, the FM benefit for the elderly was similar to that of young adults. These results were discussed in relation to speech specific auditory processing.

On the Conversant-Specificity of Stochastic Turn-Taking Models

Kornel Laskowski; Carnegie Mellon University, USA

Wed-P-19-12, Poster

Stochastic turn-taking models provide stationary estimates of the probability of a conversant’s incipient speech activity, given their own and their interlocutors’ recent speech activity. Existing research suggests that such models may be conversant specific, and even conversant-discriminative. The present work establishes this explicitly. It is shown that: (1) the conditioning context can be relaxed to exploit speech activity which need not be attributed to specific interlocutors; (2) the same duration of context can yield better results with a more statistically sound framework; and (3) results further improve asymptotically with the consideration of longer conditioning histories. The findings indicate that inter-conversant variability is a major contributor of variability across stochastic turn-taking models.
We investigated a single-ended speech intelligibility estimation method that does not require clean speech reference signal, using the features defined in the ITU-T standard P.563. We selected two sets of features from the P.563 features; the basic nine feature set, and the extended 31 feature set with 22 additional features for more accurate description of the degraded speech. Four hundred noise samples were added to speech, and about 70% of these samples were used to extract the feature sets to train a support vector regression (SVR) model. The trained models were used to estimate the intelligibility for speech degraded with the remaining 30% of unknown noise samples. The proposed method showed a root mean square error (RMSE) value of about 0.16 and correlation with subjective intelligibility of about 0.84 for speech distorted with unknown noise with either of the feature set. These results were higher than the double-sided estimation using frequency-weighted SNR calculated in critical frequency bands, which require the clean reference signal. We believe this level of accuracy proves the proposed method to be applicable to real-time speech quality monitoring in the field.

Spectral Tilt Modelling with GMMs for Intelligibility Enhancement of Narrowband Telephone Speech
Emma Jokinen, Ulpu Remes, Marko Takainen, Kalle Palomäki, Mikko Kurimo, Paavo Alku; Aalto University, Finland

In mobile communications, post-processing methods are used to improve the intelligibility of speech in adverse background noise conditions. In this study, post-processing based on modelling the Lombard effect is investigated. The study focuses on comparing different spectral envelope estimation methods together with Gaussian mixture modelling in order to change the spectral tilt of speech in a post-processing algorithm. Six spectral envelope estimation methods are compared using objective distortion measures as well as subjective word-error rate and quality tests in different near-end noise conditions. Results show that one of the envelope estimation methods, stabilised weighted linear prediction, yielded statistically significant improvement in intelligibility over unprocessed speech.

Analyzing Perceptual Dimensions of Conversational Speech Quality
Friedemann Köster, Sebastian Möller; T-Labs, Germany

Most telecommunication systems are used for communication between two people which interact during a conversation. In general, the quality of conversational speech is the major indicator for telecommunication-service providers to evaluate their systems. In this context, not only the assessment of the overall quality but also the analysis of the conversational speech quality is essential. We present an initial approach towards analyzing the conversational quality by separating a conversation in phases, and extracting individual corresponding perceptual dimensions of quality, as they are subjectively perceived by the system users. These dimensions can be combined for overall quality estimation and may separately be used to diagnose the technical reasons of quality degradation. For this reason, we review known and identify new dimensions of quality perception on the basis of subjective experiments. This enables to deeply analyze conversational speech quality for diagnosis and optimization of telecommunication systems.

Interplay of Informational Content and Energetic Masking in Speech Perception in Noise
Vincent Aubanel, Chris Davis, Jeesun Kim; University of Western Sydney, Australia

It seems plausible that different regions of the speech signal convey different amounts of information. Understanding which aspects of the signal convey information is important for understanding speech perception, particularly when this occurs in noisy environments. The so called cochlea-scaled entropy (CSE) measure is an index of spoken information based on the distribution of spectral energy over consonant/vowel time scales that is defined independently of potential noise corruption. In speech in noise, however, energetic masking distorts information, because it suppresses certain spectro-temporal regions. This study explored the interplay of informational content (defined by CSE) and energetic masking in explaining the listeners ability to understand speech in noise. Using a priming paradigm, mixtures of speech and speech-shape noise were presented to listeners in an identification task. Sentences were preceded by previews consisting of either low or high informational content. Both types yielded a similar performance increase of around 19%. Although the low information preview transmitted less target information it had a greater overlap with the more energetic regions of the target sentence (i.e., those that were less masked). This could explain why both preview types were effective and calls for a consideration of both measures in understanding speech recognition in noise.

On Spectral and Time Domain Energy Reallocation for Speech-in-Noise Intelligibility Enhancement
Tudor-Cătălin Zorilă, Yannis Stylianou; University of Crete, Greece

This paper addresses the problem of increasing speech-in-noise intelligibility under the constraint of energy preservation. Two recently proposed algorithms which have been shown to be very successful in this problem according to two large formal listening tests are reviewed and a hybrid system which combines the properties of the two methods is suggested. The first technique, which is a frequency domain approach, is re-implemented providing clarifications on its energy reallocation strategy. Based on objective measures well correlated with human perception, we show that our implementation performs similarly to the original approach. Moreover, this is combined with a dynamic range compression algorithm from the second method to allow reallocation of energy over time as well. Experiments with speech shaped noise (SSN) and competing speaker (CS) noise masks at various SNRs indicate that the hybrid system outperforms the individual algorithms in terms of intelligibility scores.
Objective Quality Evaluation of Noise-Suppressed Speech: Effects of Temporal Envelope and Fine-Structure Cues
Fei Chen1, Yi Hu2; 1University of Hong Kong, China; 2University of Wisconsin-Milwaukee, USA
Wed-P-20-6, Poster

While temporal envelope and fine-structure cues are known to be good predictors for speech intelligibility, it is not clear how well they are correlated with subjective quality ratings, particularly those using noise-suppressed speech. The present work evaluated the performance of two objective measures (i.e., NCM and TFSS), which were originally developed with primarily envelope or fine-structure cue as speech intelligibility indices, when they were applied for predicting the subjective quality ratings of noise-suppressed speech along three dimensions of signal distortion, noise distortion and overall quality. We considered a wide range of distortion introduced by four types of real-world noises at two signal-to-noise-ratio levels and by four classes of noise-suppression algorithms. This work finds that the present envelope- and fine-structure-based measures poorly predict the subjective quality ratings of noise-suppressed speech. The PESQ measure is so far the best choice in terms of objectively evaluating both subjective quality ratings and intelligibility scores of noise-suppressed speech.

Noisy Speech Enhancement Based on Long Term Harmonic Model to Improve Speech Intelligibility for Hearing Impaired Listeners
Dongmei Wang, Philipos C. Loizou, John H.L. Hansen; University of Texas at Dallas, USA
Wed-P-20-7, Poster

This study proposes a speech enhancement algorithm to improve speech intelligibility for hearing impaired listeners in adverse conditions. The proposed algorithm is based on a long term harmonic model, where the harmonics of target speech are more distinguished from noise spectrum interference. Our method consists of two stages: i) Prominent pitch estimation based on long term harmonic feature analysis and neural network classification. ii) Target speech spectrum estimation with pitch information based on long term noise spectrum extraction. The listening experiment with EAS vocoder speech shows that our algorithm is substantially beneficial for cochlear implant recipients to perceive speech in noisy environment in terms of word recognition rate.

Using Linguistic Predictability and the Lombard Effect to Increase the Intelligibility of Synthetic Speech in Noise
Cassia Valentini-Botinhao, Mirjam Wester; University of Edinburgh, UK
Wed-P-20-8, Poster

In order to predict which words in a sentence are harder to understand in noise it is necessary to consider not only audibility but also semantic or linguistic information. This paper focuses on using linguistic predictability to inform an intelligibility enhancement method that uses Lombard-adapted synthetic speech to modify low predictable words in Speech Perception in Noise (SPIN) test sentences. Word intelligibility in the presence of speech-shaped noise was measured using plain Lombard and a combination of the two synthetic voices. The findings show that the Lombard voice increases intelligibility in noise but the intelligibility gap between words in a high and low predictable context still remains. Using a Lombard voice when a word is unpredictable is a good strategy, but if a word is predictable from its context the Lombard benefit only occurs when other words in the sentence are also modified.

Speech Pre-Enhancement Using a Discriminative Microscopic Intelligibility Model
Maryam Al Dabel, Jon Barker; University of Sheffield, UK
Wed-P-20-9, Poster

We propose a new approach for optimally pre-enhancing speech signals for given noise conditions. Like others, we optimise the predicted intelligibility of the signal, however, we employ a statistical 'microscopic' intelligibility model that encodes information about which spectro-temporal speech regions are most informative. Uniquely, our optimisation strategy aims to maximise the discrimination between the correct interpretation and competing incorrect interpretations of the utterance. We present results from studies that use speech-shaped stationary noise masks and show the new strategy leads to solutions that are more varied than the simple high frequency emphasis employed in many pre-enhancement systems.

Least Squares Signal Declipping for Robust Speech Recognition
Mark J. Harvilla, Richard M. Stern; Carnegie Mellon University, USA
Wed-P-20-10, Poster

This paper introduces a novel declipping algorithm based on constrained least-squares minimization. Digital speech signals are often sampled at 16 kHz and classic declipping algorithms fail to accurately reconstruct the signal at this sampling rate due to the scarcity of reliable samples after clipping. The Constrained Blind Amplitude Reconstruction algorithm interpolates missing data points such that the resulting function is smooth while ensuring the inferred data fall in a legitimate range. The inclusion of explicit constraints helps to guide an accurate interpolation. Evaluation of declipping performance is based on automatic speech recognition word error rate and Constrained Blind Amplitude Reconstruction is shown to outperform the current state-of-the-art declipping technology under a variety of conditions. Declipping performance in additive noise is also considered.

Poster Session 21: Speech and Language Processing — General Topics

Semi-Supervised Training for Bottle-Neck Feature based DNN-HMM Hybrid Systems
Haihua Xu1, Hang Su2, Eng Siong Chng1, Haizhou Li3; 1Nanyang Technological University, Singapore; 2ICSJ, USA; 3A*STAR, Singapore
Wed-P-21-1, Poster

In this paper, we investigate semi-supervised training (SST) method in various state-of-the-art acoustic modeling techniques, using bottle-neck and corresponding tandem features. These techniques include subspace GMM, tanh-neuron deep neural network (DNN), and a generalized soft-maxout (p-norm) DNN. We demonstrate that SST may lead up to 2% Word Error Rate (WER) reduction using all these techniques in each case, and the best one comes from tandem feature based p-norm DNN system. In addition to...
Deep neural networks (DNNs) have recently become the state of the art technology in speech recognition systems. In this paper we propose a new approach to constructing large high quality unsupervised sets to train DNN models for large vocabulary speech recognition. The core of our technique consists of two steps. We first redecode speech logged by our production recognizer with a very accurate (and hence too slow for real-time usage) set of speech models to improve the quality of ground truth transcripts used for training alignments. Using confidence scores, transcript length and transcript flattening heuristics designed to cull salient utterances from three decades of speech per language, we then carefully select training data sets consisting of up to 15K hours of speech to be used to train acoustic models without any reliance on manual transcription. We show that this approach yields models with approximately 18K context dependent states that achieve 10% relative improvement in large vocabulary dictation and voice-search systems for Brazilian Portuguese, French, Italian and Russian languages.

Recent Advances in ASR Applied to an Arabic Transcription System for Al-Jazeera
Patrick Cardinal¹, Ahmed Ali², Najim Dehak¹, Yu Zhang¹, Tuka Al Hanai¹, Yifan Zhang², James R. Glass¹, Stephan Vogel²; ¹MIT, USA; ²Qatar Computing Research Institute, Qatar

This paper describes a detailed comparison of several state-of-the-art speech recognition techniques applied to a limited Arabic broadcast news dataset. The different approaches were all trained on 50 hours of transcribed audio from the Al-Jazeera news channel. The best results were obtained using i-vector-based speaker adaptation in a training scenario using the Minimum Phone Error (MPE) criteria combined with sequential Deep Neural Network (DNN) training. We report results for two different types of test data: broadcast news reports, with a best word error rate (WER) of 17.86%, and a broadcast conversations with a best WER of 29.85%. The overall WER on this test set is 25.6%.

rwth1m — The RWTH Aachen University Neural Network Language Modeling Toolkit
Martin Sundermeyer, Ralf Schlüter, Hermann Ney; RWTH Aachen University, Germany

We present a novel toolkit that implements the long short-term memory (LSTM) neural network concept for language modeling. The main goal is to provide a software which is easy to use, and which allows fast training of standard recurrent and LSTM neural network language models. The toolkit obtains state-of-the-art performance on the standard Treebank corpus. To reduce the training time, BLAS and related libraries are supported, and it is possible to evaluate multiple word sequences in parallel. In addition, arbitrary word classes can be used to speed up the computation in case of large vocabulary sizes. Finally, the software allows easy integration with SRILM, and it supports direct decoding and rescoring of HTK lattices. The toolkit is available for download under an open source license.

Language Modeling with Sum-Product Networks
Wei-Chen Cheng¹, Stanley Kok¹, Hoai Vu Pham¹, Hai Leong Chieu², Kian Ming A. Chai²; ¹SUTD, Singapore; ²DSO, Singapore

Sum product networks (SPNs) are a new class of deep probabilistic models. They can contain multiple hidden layers while keeping their inference and training times tractable. An SPN consists of interleaving layers of sum nodes and product nodes. A sum node can be interpreted as a hidden variable, and a product node can be viewed as a feature capturing rich interactions among an SPN’s inputs. We show that the ability of SPN to use hidden layers to model complex dependencies among words, and its tractable inference and learning times, make it a suitable framework for a language model. Even though SPNs have been applied to a variety of vision problems [1, 2], we are the first to use it for language modeling. Our empirical comparisons with six previous language models indicate that our SPN has superior performance.

Improving Deep Neural Network Acoustic Modeling for Audio Corpus Indexing Under the IARPA Babel Program
Xiaodong Cui¹, Brian Kingsbury¹, Jia Cui¹, Bhuvana Ramabhadran¹, Andrew Rosenberg⁷, Mohammad Sadegh Kasoob³, Owen Rambow³, Nizar Habash³, Vaibhava Goel¹; ¹IBM T.J. Watson Research Center, USA; ²CUNY Queens College, USA; ³Columbia University, USA

This paper is focused on several techniques that improve deep neural network (DNN) acoustic modeling for audio corpus indexing in the context of the IARPA Babel program. Specifically, fundamental frequency variation (FFV) and channel-aware (CA) features and data augmentation based on stochastic feature mapping (SFM) are investigated not only for improved automatic speech recognition (ASR) performance but also for their impact to the final spoken term detection on the pre-indexed audio corpus. Experimental results on development languages of Babel option period one show that the improved DNN acoustic models can reduce word error rates in ASR and also help the keyword search performance compared to already competitive DNN baseline systems.

Cross-Language Transfer of Semantic Annotation via Targeted Crowdsourcing
Shammur Absar Chowdhury¹, Arindam Ghosh¹, Evgeny A. Stepanov¹, Ali Orkan Bayer¹, Giuseppe Riccardi¹, Ioannis Klasinas²; ¹Universitá di Trento, Italy; ²Technical University of Crete, Greece

The development of a natural language speech application requires the process of semantic annotation. Moreover multilingual porting of speech applications increases the cost and complexity of the annotation task. In this paper we address the problem of transferring the semantic annotation of the source language corpus to a low-resource target language via crowdsourcing.

NOTES
current crowdsourcing approach faces several problems. First, the available crowdsourcing platforms have skewed distribution of language speakers. Second, speech applications require domain-specific knowledge. Third, the lack of reference target language annotation, makes crowdsourcing worker control very difficult. In this paper we address these issues on the task of cross-language transfer of domain-specific semantic annotation from an Italian spoken language corpus to Greek, via targeted crowdsourcing. The issue of domain knowledge transfer is addressed by priming the workers with the source language concepts. The lack of reference annotation is coped with a consensus-based annotation algorithm. The quality of annotation transfer is assessed using source language references and inter-annotator agreement. We demonstrate that the proposed computational methodology is viable and achieves acceptable annotation quality.

**Probabilistic Enrichment of Knowledge Graph Entities for Relation Detection in Conversational Understanding**

Dilek Hakkani-Tür, Asli Celikyilmaz, Larry Heck, Gokhan Tur, Geoff Zweig; Microsoft, USA

Knowledge encoded in semantic graphs such as Freebase has been shown to benefit semantic parsing and interpretation of natural language user utterances. In this paper, we propose new methods to assign weights to semantic graphs that reflect common usage types of the entities and their relations. Such statistical information can improve the disambiguation of entities in natural language utterances. Weights for entity types can be derived from the populated knowledge in the semantic graph, based on the frequency of occurrence of each type. They can also be learned from the usage frequencies in real world natural language text, such as related Wikipedia documents or user queries posed to a search engine. We compare the proposed methods with the unweighted version of the semantic knowledge graph for the relation detection task and show that all weighting methods result in better performance in comparison to using the unweighted version.

**Automatic Speech Recognition and Translation of a Swiss German Dialect: Walliserdeutsch**

Philip N. Garner, David Imseng, Thomas Meyer; Idiap Research Institute, Switzerland

Walliserdeutsch is a Swiss German dialect spoken in the south west of Switzerland. To investigate the potential of automatic speech processing of Walliserdeutsch, a small database was collected based mainly on broadcast news from a local radio station. Experiments suggest that automatic speech recognition is feasible: use of another (Swiss German) database shows that the small data size lends itself to bootstrapping from other data; use of Kullback-Leibler HMM suggests that phoneme mapping techniques can compensate for a grapheine-based dictionary. Experiments also indicate that statistical machine translation is feasible; the difficulty of small data size is offset by the close proximity to (high) German.

**Building Resources for Algerian Arabic Dialects**

S. Harrat 1, K. Meftouh 2, M. Abbas 3, K. Smalii 4, 1ENS Bouzareah, Algeria; 2UBMA, Algeria; 3CRSTDLA, Algeria; 4LORIA, France

The Algerian Arabic dialects are under-resourced languages, which lack both corpora and Natural Language Processing (NLP) tools, although they are increasingly used in written form, especially on social media and forums. We aim through this paper, and for the first time, to build parallel corpora for Algerian dialects, because our ultimate purpose is to achieve a Machine Translation (MT) for Modern Standard Arabic (MSA) and Algerian dialects (AD), in both directions. We also propose language tools to process these dialects. First, we developed a morphological analysis model of dialects by adapting BAMA, a well-known MSA analyzer. Then we propose a diacritization system, based on a MT process which allows to restore the vowels to dialects corpora. And finally, we propose results on machine translation between MSA and Algerian dialects.

**Show and Tell Session 2**

Garnet Foyer
13:30 – 15:30, Wednesday 17 September 2014
Chair: Alvina Goh Siew Wee

**An Educational Platform to Capture, Visualize and Analyze Rare Singing**

P. Chawah 1, S.K. Al Kork 2, T. Fux 1, Martine Adda-Decker 1, A. Amelot 1, N. Audibert 1, B. Denby 2, G. Dreyfus 1, A. Jaumard-Hakoun 1, C. Pillot-Loiseau 1, P. Roussel 1, M. Stone 1, Kele Xu 1, L. Crevier-Buchman 1; 1LPP (UMR 7018), France; 2UPMC, France; 3SIGMA (UMR 7084), France; 4University of Maryland, USA

This paper presents an early version of an open extendable research and educational platform to support users in learning and mastering the different types of rare-singing. The platform is interfaced with a portable helmet to synchronously capture multiple signals during singing in a non-laboratory environment. Collected signals reflect articulatory movements and induced vibrations. The platform consists of four main modules: i) a capture and recording module, ii) a data replay (post processing) module, iii) an acoustic auto adaptation learning module, iv) and a 3D visualization sensory motor learning module. Our demo will focus on the first two modules. The system has been tested on two rare endangered singing musical styles, the Corsican “Cantu in Paghjella”, and the Byzantine hymns from Mount Athos, Greece. The versatility of the approach is further demonstrated by capturing a contemporary singing style known as “Human Beat Box.”

**Single-Channel Speech Enhancement Based on Non-Negative Matrix Factorization and Online Noise Adaptation**

Kwang Myung Jeon 1, Chan Jun Chun 1, Woo Kyeong Seong 1, Hong Kook Kim 1, Myung Kyu Choi 2; 1GIST, Korea; 2Samsung Electronics, Korea

In this paper, we demonstrate a simulator for real-time speech enhancement based on a non-negative matrix factorization (NMF) technique. In particular, we propose an online noise adaptation method in an NMF framework, which is activated during non-speech intervals and used for adapting noise bases for NMF. Thus, incoming noisy speech is decomposed by using such adapted noise bases and universal speech bases that can be developed through training with examples of speech data. It is shown from the experiments that the proposed method improves speech separation performance and perceptual speech quality.

**Notes**
Intelligibility of High-Pitched Vowel Sounds in the Singing and Speaking of a Female Cantonese Opera Singer

Dieter Maurer¹, Peggy Mok², Daniel Friedrichs³, Volker Dellwo³; ¹Zürcher Hochschule der Künste, Switzerland; ²Chinese University of Hong Kong, China; ³Universität Zürich, Switzerland

The question whether or not vowel quality can be maintained at F0 of the sounds exceeding statistical F1 of “normal” speech is still a matter of debate. The present study investigates the perception of long Cantonese vowels /i, y, oe, a, u/ spoken and sung in (C)V and (C)V:S context by a well-known female Cantonese Opera singer in the range of F0 of c. 560–860Hz. 172 high-pitched syllables or isolated vowel sounds were selected and extracted from a live recording. Corresponding vowel perception was tested in a listening test performed by 26 students of linguistics. All six vowels proved to be identifiable > 80% up to F0 of c. 700Hz, and sounds of /i, a, u/ proved to be identifiable > 80% up to a range of F0 of c. 820–860Hz. Confusion matrices are provided in the present paper and spectral illustrations of all sounds are presented at http://is2014.phones-and-phonemes.org.

Iterative Refinement of Amplitude and Phase in Single-Channel Speech Enhancement

Pejman Mowlaee¹, Mario Kaoru Watanabe¹, Rahim Saeidi²; ¹Technische Universität Graz, Austria; ²University of Eastern Finland, Finland

While the state-of-the-art speech enhancement methods are focused on the modification of the noisy spectral amplitude, our recent findings demonstrate positive impact of incorporating the speech phase spectrum in speech enhancement. In this show and tell proposal, we demonstrate the recent progress towards utilizing the phase information in closed-loop iterative manner leading to the joint enhancement of amplitude and phase spectra. In this show and tell, we provide a demo where a recorded speech corrupted with noise is enhanced by an iterative refinement of amplitude and phase. The improvement in speech enhancement in various signal-to-noise ratios is justified by improved perceived speech quality as well as not degrading the speech intelligibility performance.

eLite-HTS: A NLP Tool for French HMM-Based Speech Synthesis

Sophie Roekhaut¹, Sandrine Brogaux¹, Richard Beaufort², Thierry Dutoit²; ¹Université catholique de Louvain, Belgium; ²Nuance Communications, Belgium;

This paper presents eLite-HTS, a web service which generates input files for the training and synthesis stages of a French HMM-based synthesizer using the HTS toolkit.

SARA — Singapore’s Automated Responsive Assistant for the Touristic Domain

Andrea I. Niculescu, Rafael E. Banchs, Ridsong Jiang, Seokhwan Kim, Kheng Hui Yeo, Arthur Niswar; A*STAR, Singapore

In this paper we describe SARA — Singapore’s Automated Responsive Assistance — an Android mobile phone application for touristic information in Singapore. The application provides information about local attractions, restaurants, sightseeing direction and transportation services. Using a GPS integrated module SARA is able to detect the user’s location on a map providing orientation and direction help. Users can use speech, text or scanned QR code to interact with SARA. Input/output modalities for SARA include natural language in form of text or speech. A short video about the main features of our Android application can be seen here: http://vimeo.com/91620644. Currently, SARA supports only English, but we are working towards a multilingual support. For test purposes we created a web version of SARA that can be tested for Chinese and English text input at: http://iris12.astar.edu.sg/StatTour/.

The Speech Recognition Virtual Kitchen: Launch Party

Andrew Plummer¹, Eric Riebling², Anuj Kumar², Florian Metze², Eric Fosler-Lussier¹, Rebecca Bates³; ¹Ohio State University, USA; ²Carnegie Mellon University, USA; ³Minnesota State University, USA

We present updates to the Speech Recognition Virtual Kitchen (SRVK) environment, a repository of pre-configured Virtual Machines (VMs) containing tools and experiments in the speech and language field. SRVK promotes community sharing of research techniques, fosters innovative experimentation, and provides solid reference systems as a tool for education, research, and evaluation. VMs provide a consistent environment for experimentation, without requiring tedious installation of many individual tools, a web-based community platform complements the VMs, allowing users to jointly explore, learn and collaborate using VMs. In this Show&Tell demo, we present the infrastructure to the speech community, along with several example VMs and a set of online error analysis tools. We solicit feedback from the community, in order to further guide development of the kitchen, which we hope to grow into a widely used community resource.

System for Automated Speech and Language Analysis (SALSA)

Kyle Marek-Spartz, Benjamin Knoll, Robert Bill, Thomas Christie, Serguei Pakhomov; University of Minnesota, USA

SALSA automates cognitive test administration and scoring. These tests are used to characterize cognitive impairment resulting from neurodegenerative disease, traumatic brain injury, and drug toxicity; however, they are currently performed manually, limiting their utility for large clinical populations and longitudinal assessments. We present a fully automated, comprehensive system for collecting spoken test responses with mobile and telephony platforms, using an open-source automatic speech recognition engine (KALDI) to calculate a range of speech characteristics that may be useful in assessment of cognitive function.

Pronunciation Practice Support System for Children Who Have Difficulty Correctly Pronouncing Words

Ikuyo Masuda-Katsuse; Kindai University, Japan

We developed a system with which children who have difficulty correctly pronouncing words to practice their pronunciation. It allows exercises to be individually tailored to each child’s pronun-
We, Institute for Infocomm Research (I2R), in Singapore, have developed a technology that allows for the creation of high-quality natural singing. Singing well is not trivial. Professional singers are trained and have developed skills in rhythm, precision in pitch, vocal control and personal styling, etc. to produce appealing singing. However, high-quality natural singing is appealing to everyone. Given an input singing, our technology will convert it into a nice piece of singing, perfecting it with correct rhythm and fundamental frequency over the time. Now, you can simply read the lyrics, record in our mobile apps and generate a pleasant and personalized singing with your own vocal characteristics.

Automated Production of True-Cased Punctuated Subtitles for Weather and News Broadcasts

Joris Driesen, Alexandra Birch, Simon Grimsey, Saeid Safarfashandi, Juliet Gauthier, Matt Simpson, Steve Renals, University of Edinburgh, UK; Red Bee Media, UK

Providing subtitling for multimedia content is a highly costly process. Any system aimed at automating at least part of this process may therefore yield significant economic benefits for content providers. In this paper, we present an integrated automatic system capable of automatically subtitling weather forecasts and news broadcasts. In this system, a number of different modules are stringed together, each performing a single processing step in the pipeline. An ASR (Automatic Speech Recognition) module first converts raw audio into an uninterrupted stream of written words. A decision tree classifier then marks sentence boundaries in the resulting word sequence. Finally, a SMT (Statistical Machine Translation) module ‘translates’ the resulting sentences into punctuated true-cased text. The system has been developed in close cooperation with Red Bee Media and will be deployed in their commercial production pipeline.

F* Speech2Singing Perfects Everyone’s Singing

Minghui Dong, S.W. Lee, Haizhou Li, Paul Chan, Xuejian Peng, Jochen Walter Ehnes, Dongyan Huang, A*STAR, Singapore

High-quality natural singing is appealing to everyone. However, singing well is not trivial. Professional singers are trained and practised over the years for various singing skills, such as sense of rhythm, precision in pitch, vocal control and personal styling etc. We, Institute for Infocomm Research (I2R), in Singapore, have developed a personalized, real-time singing synthesis technology for the general public. Given an input singing (maybe in mediocre to poor quality), our technology will convert it into a nice piece of singing, perfecting it with correct rhythm and fundamental frequency over the time. Now, you can simply read the lyrics, record in our mobile apps and generate a pleasant and personalized singing with your own vocal characteristics.

Oral Session 29: Language, Dialect and Accent Recognition

Garnet 213-218
16:00 – 18:00, Wednesday 17 September 2014
Chairs: John Hansen and Ville Hautamäki

Spoken Language Recognition Based on Senone Posterior

Luciana Ferrer, Yun Lei, Mitchell McLaren, Nicolas Scheffer; SRI International, USA

This paper explores in depth a recently proposed approach to spoken language recognition based on the estimated posteriors for a set of senones representing the phonetic space of one or more languages. A neural network (NN) is trained to estimate the posterior probabilities for the senones at a frame level. A feature vector is then derived for every sample using these posteriors. The effect of the language used in training the NN and the number of senones are studied. Speech-activity detection (SAD) and dimensionality reduction approaches are also explored and Gaussian and NN backends are compared. Results are presented on heavily degraded speech data. The proposed system is shown to give over 40% relative gain compared to a state-of-the-art language recognition system at sample durations from 3 to 120 seconds.

Automatic Language Identification Using Long Short-Term Memory Recurrent Neural Networks

Javier Gonzalez-Dominguez, Ignacio Lopez-Moreno, Hasim Sak, Joaquin Gonzalez-Rodriguez, Pedro J. Moreno, Google, USA; Universidad Autónoma de Madrid, Spain

This work explores the use of Long Short-Term Memory (LSTM) recurrent neural networks (RNNs) for automatic language identification (LID). The use of RNNs is motivated by their better ability in modeling sequences with respect to feed forward networks used in previous works. We show that LSTM RNNs can effectively exploit temporal dependencies in acoustic data, learning relevant features for language-discrimination purposes. The proposed approach is compared to baseline i-vector and feed forward Deep Neural Network (DNN) systems in the NIST Language Recognition Evaluation 2009 dataset. We show LSTM RNNs achieve better performance than our best DNN system with an order of magnitude fewer parameters. Further, the combination of the different systems leads to significant performance improvements (up to 28%).

Robust Language Recognition via Adaptive Language Factor Extraction

Brecht Desplanques, Kris Demuynck, Jean-Pierre Martens; Ghent University, Belgium

This paper presents a technique to adapt an acoustically based language classifier to the background conditions and speaker accents. This adaptation improves language classification on a broad spectrum of TV broadcasts. The core of the system consists of an i-vector-based setup in which language and channel variabilities are modeled separately. The subsequent language classifier (the backend) operates on the language factors, i.e. those features in the extracted i-vectors that explain the observed language variability. The proposed technique adapts the language variability model to
the background conditions and to the speaker accents present in
the audio. The effect of the adaptation is evaluated on a 28 hours
corpus composed of documentaries and monolingual as well as
multilingual broadcast news shows. Consistent improvements in
the automatic identification of Flemish (Belgian Dutch), English
and French are demonstrated for all broadcast types.

Dialect Levelling in Finnish: A Universal Speech
Attribute Approach
Harmin Behuvan 1, Ville Hautamäki 1, Sabato Marco
Siniscalchi 2, Elie Khoury 3, Tommi Kurki 4, Tomi
Kinnunen 1, Chin-Hui Lee 5; 1University of Eastern
Finland, Finland; 2Università Kore di Enna, Italy; 3Idiap
Research Institute, Switzerland; 4University of Turku,
Finland; 5Georgia Institute of Technology, USA

We adopt automatic language recognition methods to study di-
alect levelling — a phenomenon that leads to reduced structural
differences among dialects in a given spoken language. In terms
of dialect characterisation, levelling is a nuisance variable that
adversely affects recognition accuracy: the more similar two
dialects are, the harder it is to set them apart. We address levelling
in Finnish regional dialects using a new SAPU (Satakunta in Speech)
corpus containing material from Satakunta (South-Western Finland)
between 2007 and 2013. To define a compact and universal set of
sound units to characterize dialects, we adopt speech attributes
features, namely manner and place of articulation. It will be shown
that speech attribute distributions can indeed characterise differ-
ences among dialects. Experiments with an i-vector system suggest
that (1) the attribute features achieve higher dialect recognition
accuracy and (2) they are less sensitive against age-related levelling
in comparison to traditional spectral approach.

Improving Native Accent Identification Using Deep
Neural Networks
Mingming Chen, Zhanlei Yang, Hao Zheng, Wenju Liu;
Chinese Academy of Sciences, China

We adopt automatic language recognition methods to study di-
alect levelling — a phenomenon that leads to reduced structural
differences among dialects in a given spoken language. In terms
of dialect characterisation, levelling is a nuisance variable that
adversely affects recognition accuracy: the more similar two
dialects are, the harder it is to set them apart. We address levelling
in Finnish regional dialects using a new SAPU (Satakunta in Speech)
corpus containing material from Satakunta (South-Western Finland)
between 2007 and 2013. To define a compact and universal set of
sound units to characterize dialects, we adopt speech attributes
features, namely manner and place of articulation. It will be shown
that speech attribute distributions can indeed characterise differ-
ces among dialects. Experiments with an i-vector system suggest
that (1) the attribute features achieve higher dialect recognition
accuracy and (2) they are less sensitive against age-related levelling
in comparison to traditional spectral approach.

Notes

Foreign Accent Recognition Based on Temporal
Information Contained in Lowpass-Filtered Speech
Marie-José Kolly, Adrian Leemann, Volker Dellwo;
Universität Zürich, Switzerland

Can the foreign accent of a speaker be recognized based on
suprasegmental temporal information? For a perception experi-
ment we created stimuli based on German sentences read by six
French and six English speakers. These foreign-accented sentences
were manipulated by (1) applying a lowpass filter with a cutoff
frequency of 300 Hz and (2) applying the same lowpass filter and
monotonizing F0. In a between-subject 2AFC perception experi-
ment we tested the accent recognition ability of 15 Swiss German
listeners per signal manipulation condition. The results showed
that speakers' native language could be recognized above chance
in both conditions. However, listeners obtained significantly lower
recognition scores in the monotonized condition. Furthermore,
higher recognition scores were obtained for French-accented
speech in the monotonized condition, a result that is discussed in
light of research on speech rhythm. We further report an effect
for speaker within each accent group. The results suggest that
suprasegmental temporal information allows for foreign accent
recognition to some degree.

Oral Session 30: Adaptation
Peridot 202-203
16:00 – 18:00, Wednesday 17 September 2014
Chairs: Florian Metze and Mike Seltzer

Adaptation of Deep Neural Network Acoustic
Models Using Factorised I-Vectors
Penny Karanasou, Yongqiang Wang, Mark J.F. Gales,
Philip C. Woodland; University of Cambridge, UK

In this paper, we utilize deep neural networks (DNNs) to auto-
matically identify native accents in English and Mandarin when
no text, speaker or gender information is available for the speech
data. Compared to the Gaussian mixture model (GMM) based
conventional methods, the proposed method benefits from two
main advantages: first, DNNs are discriminative models which can
provide better discrimination on confusion regions of different
accents; second, they have the hierarchical nonlinear feature
extraction capability which can learn discriminative high-level
features for the specified task. In detail, the speech data of all
accents is used to train DNNs, and in the testing stage, we first
identify the accent label of each frame, then determine the sentence
label by the majority voting conducted on the frame labels. The
experiments on accented English and Mandarin corpus demon-
strate that, compared to the GMM based methods, our proposed
method can significantly improve the frame accuracy as well as
sentence accuracy on the test set. Moreover, the performance of
the proposed method can be further improved by using context
information.

Regularized Feature-Space Discriminative
Adaptation for Robust ASR
Takashi Fukuda 1, Osamu Ichikawa 1, Masafumi
Nishimura 1, Steven J. Rennie 2, Vaibhava Goel 2; 1IBM
Research Tokyo, Japan; 2IBM T.J. Watson Research
Center, USA

In this paper, we utilize deep neural networks (DNNs) to auto-
matically identify native accents in English and Mandarin when
no text, speaker or gender information is available for the speech
data. Compared to the Gaussian mixture model (GMM) based
conventional methods, the proposed method benefits from two
main advantages: first, DNNs are discriminative models which can
provide better discrimination on confusion regions of different
accents; second, they have the hierarchical nonlinear feature
extraction capability which can learn discriminative high-level
features for the specified task. In detail, the speech data of all
accents is used to train DNNs, and in the testing stage, we first
identify the accent label of each frame, then determine the sentence
label by the majority voting conducted on the frame labels. The
experiments on accented English and Mandarin corpus demon-
strate that, compared to the GMM based methods, our proposed
method can significantly improve the frame accuracy as well as
sentence accuracy on the test set. Moreover, the performance of
the proposed method can be further improved by using context
information.
often used for porting old acoustic models into new domains. Discriminative schemes for model adaptation based on MMI and MPE objective functions are also utilized. For feature-space adaptations, one extension to the well-known feature-space discriminative training (FMPE) algorithm, feature-space discriminative adaptation, was recently proposed to adapt FMPE transforms. Feature-space discriminative adaptation was shown to work well for some situations when sufficient adaptation data is available.

This paper improves the feature-space discriminative adaptation by introducing a regularization term for an indirect differential computation of the FMPE objective function, and also by updating the acoustic models with MAP instead of ML criterion during the adaptation. The proposed method performed favorably for the adaptation conditions from general-purpose LVCSR to automotive environments with small amounts of adaptation data, and yielded 4.4% relative improvement as compared with MAP-adapted system without using the FMPE adaptation.

Towards Speaker Adaptive Training of Deep Neural Network Acoustic Models

Yajie Miao, Hao Zhang, Florian Metze; Carnegie Mellon University, USA

Wed-O-30-3

We investigate the concept of speaker adaptive training (SAT) in the context of deep neural network (DNN) acoustic models. Previous studies have shown success of performing speaker adaptation for DNNs in speech recognition. In this paper, we apply SAT to DNNs by learning two types of feature mapping neural networks. Given an initial DNN model, these networks take speaker i-vectors as additional information and project DNN inputs into a speaker-normalized space. The final SAT model is obtained by updating the canonical DNN in the normalized feature space. Experiments on a Switchboard 110-hour setup show that compared with the baseline DNN, the SAT-DNN model brings 7.5% and 6.0% relative improvement when DNN inputs are speaker-independent and speaker-adapted features respectively. Further evaluations on the more challenging BABEL datasets reveal significant word error rate reduction achieved by SAT-DNN.

Component Structuring and Trajectory Modeling for Speech Recognition

Arseniy Gorin, Denis Jouvet; LORIA, France

Wed-O-30-4

When the speech data are produced by speakers of different age and gender, the acoustic variability of any given phonetic unit becomes large, which degrades speech recognition performance. A way to go beyond the conventional Hidden Markov Model is to explicitly include speaker class information in the modeling. Speaker classes can be obtained by unsupervised clustering of the speech utterances. This paper introduces a structuring of the Gaussian components of the GMM densities with respect to speaker classes. In a first approach, the structuring of the Gaussian components is combined with speaker class-dependent mixture weights. In a second approach, the structuring is used with mixture transition matrices, which add dependencies between Gaussian components of mixture densities (as in straddled GMMs). The different approaches are evaluated and compared in detail on the TIDIGITS task. Significant improvements are obtained using the proposed approaches based on structured components. Additional results are reported for phonetic decoding on the NELOGOS database, a large corpus of French telephone data.
A Sparse Reconstruction Method for Speech Source Localization Using Partial Dictionaries Over a Spherical Microphone Array

Kushagra Singhal, Rajesh M. Hegde; IIT Kanpur, India

Wed-O-31-1

Sparse reconstruction methods have been used extensively for source localization over uniform linear arrays and circular arrays. In this paper a sparse reconstruction method for speech source localization using partial dictionaries over a spherical microphone array is proposed. The source localization method proposed in this work addresses two important research issues. It formulates the source localization problem in the spherical harmonics domain as a sparse reconstruction problem. Subsequently, a low complexity method to estimate the direction of arrival (DOA) of multiple sources is also proposed by using partial elevation angle dictionaries. The use of such dictionaries reduces the complexity of the search involved in the two dimensional DOA estimation. Source localization experiments are conducted at different SNRs and compared with conventional DOA estimation methods like MUSIC and MVDR. The experimental results obtained from the proposed method indicate a reasonable reduction in the localization error.

A Robust TDOA Estimation Method for In-Car-Noise Environments

Weiwei Cui, Jaeyeon Cho, Seungyeol Lee; Samsung Electronics, Korea

Wed-O-31-2

Time difference of arrival (TDOA) estimation is one of the key techniques in array signal processing and has wide applications in hands-free speech interface. TDOA is used to derive the time difference of signal propagation from source to the spatially separated microphones. The acoustic transfer functions ratio (ATF-s ratio) method is one of the robust approaches to get TDOA estimate in presence of strong background noise and in reverberant environment. Although the effectiveness is proved by various simulations in related studies, the anti-noise or anti-reverberation capability of ATF-s ratio is limited in real acoustic environment. In this work, a frequency domain partial-whitening ATF-s ratio method is proposed for detecting the time delay in an in-car mobile system. To test the efficacy of the proposed method, the experiments were conducted in a car running at a speed of about 100km/h on the highway, and the results show that the percentage of TDOA estimation anomaly is reduced by 15%~25%.

Robust Low-Resource Sound Localization in Correlated Noise

Lorin Netsch, Jacek Stachurski; Texas Instruments, USA

Wed-O-31-3

In this paper we address the problem of sound source location using the time difference of arrival (TDOA) technique in an environment containing stationary correlated noise. We present a robust low-complexity method for enhancing estimation of sound direction, augmenting the well-known Generalized Cross-Correlation with Phase Transform (GCC-PHAT) approach. In the proposed method, the estimated cross-spectrum of a correlated background noise is subtracted from the observed spectrum. This effectively removes the phase distortion introduced by the interfering noise and significantly improves the robustness of the sound direction estimate. We test the performance of this approach on data collected and processed with a low-resource embedded platform. Results illustrate substantially enhanced performance over the baseline GCC-PHAT sound localization.

Direction-of-Arrival Estimation of Multiple Speakers Using a Planar Array

Dongwen Ying, Ruohua Zhou, Junfeng Li, Jielin Pan, Yonghong Yan; Chinese Academy of Sciences, China

Wed-O-31-4

This paper presents a novel method to estimate the direction-of-arrivals (DOAs) of multiple speakers based on time-frequency sparsity. The acoustic interferences are first suppressed, and then a concave cost function (CCF) is utilized to estimate bin-wise DOAs. The DOAs of sources are subsequently identified by picking peaks in a histogram of bin-wise azimuths. Three aspects distinguish this method from conventional ones. First, a closed-form solution to bin-wise DOA is given by CCF, which replaces the extensively used grid search and enables high computational efficiency. Second, signal enhancement is employed to suppress acoustic interferences on phase spectra. Last, the time-delay weights mitigate the effect of delay outliers. The proposed method is compared with well-established methods in simulated environments. The experimental results confirmed its superiority in computational efficiency and acoustic robustness.

Weighted Spatial Bispectrum Correlation Matrix for DOA Estimation in the Presence of Interferences

Wei Xue, Shan Liang, Wenju Liu; Chinese Academy of Sciences, China

Wed-O-31-5

Besides the undirected environmental noise, the surrounding interference also brings great challenges to the robust DOA estimation of the speech source. As conventional DOA estimation methods always assume an undirected noise model, they usually cannot perform reliably when the strong interference exists. In this paper, we propose a novel interference robust DOA estimation method, which is based on the "weighted spatial bispectrum correlation matrix (WSBCM)". The WSBCM contains the spatial correlation information of bispectrum phase difference (BPD), and a new DOA estimator is further derived based on the eigenvalue analysis of the WSBCM. By formulating with WSBCM, the proposed method benefits from the redundant DOA-related information provided by the BPD. In addition, the WSBCM enables bispectrum weighting to highlight the pure speech units in the bispectrum, which further helps to improve the performance. In order to compute the bispectrum weights, a decision-directed approach is derived. The effectiveness of the proposed method is demonstrated by experiments conducted under various kinds of interference-existing scenarios.

Multi-Sources Separation for Sound Source Localization

Mariem Bouaif, Zied Lachiri; El-Manar University, Tunisia

Wed-O-31-6

Spatial localization given measurements of the sound field at a microphone array is a computationally challenging task. While big progress has been achieved in Azimuth localization, little attention has been paid to localization in distance. This paper presents a multi-sources localization model based on Time Delay of Arrival (TDOA) using only two microphones. The proposed method requires estimation of mixing parameters, which are informative.
about speakers’ positions. The mixing parameters are estimated by determining the slope of a scatter plot of the separated sound source from the time delayed mixed speech captured by each microphone. In the determined context, experiments on BSS-eval dataset show that the proposed model achieves distance accuracy of more than 92% with a less than 1.8m distance localization error.

Oral Session 32: Speech Analysis II
Peridot 201
16:00 – 18:00, Wednesday 17 September 2014
Chairs: Thomas Quatieri and Yannis Stylianou

Relating Automatic Vowel Space Estimates to Talker Intelligibility
Yi Luan, Richard Wright, Mari Ostendorf, Gina-Anne Levow; University of Washington, USA

Differences in pronunciation have been shown to underlie significant talker-dependent intelligibility differences. There are several dimensions of variability that are correlated with talker intelligibility including pitch range, vowel-space expansion, and rhythmic patterns. Prior work has shown that some of the better predictors of individual intelligibility are based on the talker's F1 by F2 vowel space, but findings are based on hand-corrected measurements on carefully balanced sets of vowels, making large scale analysis impractical. This paper proposes a novel method for automatic estimation of a talker’s vowel space using sparse expanded vowel space representations, including an approximate convex hull sampling, which are projected to a low dimensional space for intelligibility scoring. Both supervised and unsupervised mappings are used to generate an intelligibility score. Automatic intelligibility rankings are assessed in terms of correlation with an intelligibility score based on human transcription accuracy. We find that including a larger sample of vowels (beyond point vowels) leads to improved performance, obtaining correlations of roughly 0.6 for this feature alone, which is a strong result given that there are other factors that may also contribute to a talker’s intelligibility in addition to a talker’s vowel space area.

Excitation Source Analysis for High-Quality Speech Manipulation Systems Based on an Interference-Free Representation of Group Delay with Minimum Phase Response Compensation
Hideki Kawahara 1, Masanori Morise 2, Tomoki Toda 3, Hideki Banno 4, Ryuichi Nisimura 1, Toshio Irino 5; 1Wakayama University, Japan; 2University of Yamanashi, Japan; 3NAIST, Japan; 4Meijo University, Japan

A group delay-based excitation source analysis and design method is introduced for extension of TANDEM-STRAIGHT, a speech analysis, modification and synthesis system. This introduction makes all components of the system be based on interference-free representations. They are power spectrum, instantaneous frequency and group delay representations. This unification has potential to solve the major weak point of VOCODER architecture for high-quality speech manipulation applications.

Sparse Time-Frequency Representation of Speech by the Vandermonde Transform
Christian Fischer Pedersen, Tom Bäckström; FAU Erlangen-Nürnberg, Germany

Efficient speech signal representations are prerequisite for efficient speech processing algorithms. The Vandermonde transform is a recently introduced time-frequency transform which provides a sparse and uncorrelated speech signal representation. In contrast, the Fourier transform only decorrelates the signal approximately. To achieve complete decorrelation, the Vandermonde transform is signal adaptive like the Karhunen-Loève transform. Unlike the Karhunen-Loève, however, the Vandermonde transform is a time-frequency transform where the transform domain components correspond to frequency components of the analysis window. In this paper we analyze the performance of sparse speech signal representation by the Vandermonde transform. This is done by applying matching pursuit and comparing with sparse representations based on dictionaries with Fourier, Cosine, Gabor and Karhunen-Loève atoms. Our results show that Karhunen-Loève yields the best sparse signal recovery; however, this is not strictly a time-frequency transform. Of the true time-frequency transforms, Vandermonde is the most efficient for sparse speech signal representation.

Analysis and Identification of Human Scream: Implications for Speaker Recognition
Mahesh Kumar Nandwana, John H.L. Hansen; University of Texas at Dallas, USA

In this paper, we present analysis of the characteristics of scream to identify its discriminating features from neutral speech. The impact of screaming on the performance of text independent speaker recognition systems has also been reported. We have observed that speaker recognition systems are not reliable when tested with scream. Also perceptual listeners test reveal that the speaker content in scream is very less for human to distinguish and classify it. This analysis will be useful for development of robust speaker recognition systems and their implementation in real-time situations.

F0 Estimation in Noisy Speech Based on Long-Term Harmonic Feature Analysis Combined with Neural Network Classification
Dongmei Wang, Philipos C. Loizou, John H.L. Hansen; University of Texas at Dallas, USA

In this study, we propose a frequency domain F0 estimation approach based on long term Harmonic Feature Analysis combined with artificial neural network (ANN) classification. Long term spectrum analysis is proposed to gain better harmonic resolution, which reduces the spectrum interference between speech and noise. Next pitch candidates are extracted for each frame from the long term spectrum. Five specific features related to harmonic structure are computed for each candidate and combined together as a feature vector to indicate the status of each candidate. An ANN is trained to model the relation between the harmonic features and the true pitch values. In the test phase, target pitch is selected from the candidates according to the maximum output score from the ANN. Finally, post-processing is applied based on average segmental output to eliminate inconsistent or fluctuating decision errors. Experimental results show that the proposed algorithm outperforms several state-of-the-art methods for F0 estimation.
under adverse conditions, including white noise and multi-speaker babble noise.

The Influence of Pitch and Noise on the Discriminability of Filterbank Features
Malcolm Slaney, Michael L. Seltzer; Microsoft, USA
Wed-0326

Most features used for speech recognition are derived from the output of a filterbank inspired by the auditory system. The two most commonly used filter shapes are the triangular filters used in MFCC (mel-frequency cepstral coefficients) and the gammatone filters that model psychoacoustic critical bands. However, for both of these filterbanks there are free parameters that must be chosen by the system designer. In this paper, we explore the effect that different parameter settings have on the discriminability of speech sound classes. Specifically, we focus our attention on two primary parameters: the filter shape (triangular or gammatone) and the filter bandwidth. We use variations in the noise level and the pitch to explore the behavior of different filterbanks. We use the Fisher linear discriminant to give us insight about why some filterbanks perform better than others. We observe three things: 1) there are significant differences even among different implementations of the same filterbank, 2) wider filters help remove the non-informative pitch information, and 3) the Fisher criteria helps us understand why. We validate the Fisher measure with speech recognition experiments on the Aurora-4 speech corpus.

Special Session 6(b): Deep Neural Networks for Speech Generation and Synthesis II
Peridot 206
16:00 – 18:00, Wednesday 17 September 2014
Chairs: Frank Soong and Yao Qian

Prosody Contour Prediction with Long Short-Term Memory, Bi-Directional, Deep Recurrent Neural Networks
Raul Fernandez¹, Asaf Rendel², Bhuvana Ramabhadran¹, Ron Hoory²; ¹IBM T.J. Watson Research Center, USA; ²IBM Research Haifa, Israel
Wed-SP6b-3

Deep Neural Networks (DNNs) have been shown to provide state-of-the-art performance over other baseline models in the task of predicting prosodic targets from text in speech-synthesis systems. However, prosody prediction can be affected by an interaction of short- and long-term contextual factors that a static model that depends on a fixed-size context window can fail to properly capture. In this work, we look at a recurrent formulation of neural networks (RNNs) that are deep in time and can store state information from an arbitrarily large input history when making a prediction. We show that RNNs provide improved performance over DNNs of comparable size in terms of various objective metrics for a variety of prosodic streams (notably, a relative reduction of about 6% in F0 mean-square error accompanied by a relative increase of about 14% in F0 variance), as well as in terms of perceptual quality assessed through mean-opinion-score listening tests.

Modeling DCT Parameterized F0 Trajectory at Intonation Phrase Level with DNN or Decision Tree
Xiang Yin¹, Ming Lei², Yao Qian², Frank K. Soong², Lei He², Zhen-Hua Ling¹, Li-Rong Dai¹; ¹USTC, China; ²Microsoft, China
Wed-SP6b-2

In the conventional HMM-based TTS, the micro structure of F0 contour is modeled at the state level via a (clustered) decision tree. However, the decision tree based state-level modeling is difficult to capture the long term structure of speech prosody, say at intonation phrase level, due to its greedy search nature and usually sparse training data for covering a large, combinatorial number of usually long prosodic contexts in a phrase or sentence. In this study, we adopt a finite number of Discrete Cosine Transform (DCT) coefficients to capture the smoothed trend of F0 patterns of intonation phrases and then normalize the variable duration effects in phrase length. We then use DCT smoothed contours to model phrase intonations with a decision tree or a deep neural network (DNN). The remaining details or the residual F0 is then accommodated by training a state-level model in a Hierarchical Prosody Model (HPM) framework. The internal phrase models are then used to predict the intonation phrase F0 contours and then combine it with the predicted state-level F0 residuals to predict final F0 contours. Either the decision tree based or the DNN based F0 predictors, when working together with the state-level F0 residual predictors, outperform the standard, state-level HMM F0 models.

High-Order Sequence Modeling Using Speaker-Dependent Recurrent Temporal Restricted Boltzmann Machines for Voice Conversion
Toru Nakashika, Tetsuya Takiguchi, Yasuo Ariki; Kobe University, Japan
Wed-SP6b-3

This paper presents a voice conversion (VC) method that utilizes recently proposed recurrent temporal restricted Boltzmann machines (RTRBMs) for each speaker, with the goal of capturing high-order temporal dependencies in an acoustic sequence. Our algorithm starts from the separate training of two RTRBMs for a source and target speaker using speaker-dependent training data. Since each RTRBM attempts to discover abstractions at each time step, as well as the temporal dependencies in the training data, we expect that the models represent the speaker-specific latent features in the high-order spaces. In our approach, we run conversion from such speaker-specific-emphasized features of the source speaker to those of the target speaker using a neural network (NN), so that the entire network (the two RTRBMs and the NN) forms a deep recurrent neural network and can be fine-tuned. Through VC experiments, we confirmed the high performance of our method especially in terms of objective criteria in comparison to conventional VC methods such as Gaussian mixture model (GMM)-based approaches.

Sequence Error (SE) Minimization Training of Neural Network for Voice Conversion
Feng-Long Xie¹, Yao Qian², Yuchen Fan², Frank K. Soong², Haifeng Li¹; ¹Harbin Institute of Technology, China; ²Microsoft, China
Wed-SP6b-4

Neural network (NN) based voice conversion, which employs a nonlinear function to map the features from a source to a target speaker, has been shown to outperform GMM-based voice con-
In the context of speech therapy based on visual biofeedback, images of the tongue of another speaker. This work is developed on an articulatory tongue model of a reference speaker from ultrasound images.

This paper presents a method for automatically animating the articulatory tongue model. The user’s tongue movements are captured using ultrasound imaging and parameterized using the PCA-based EigenTongue technique. Extracted features are then converted into control parameters of the articulatory tongue model using Gaussian Mixture Regression. This procedure was evaluated by decoding the converted tongue movements at the phonetic level using an HMM-based decoder trained on the reference speaker’s articulatory data. Decoding errors were then manually reassessed in order to take into account possible phonetic idiosyncrasies.

Articulatory Controllable Speech Modification Based on Statistical Feature Mapping with Gaussian Mixture Models

Patrick Lumban Tobing 1, Tomoki Toda 1, Graham Neubig 1, Sakriani Sakti 1, Satoshi Nakamura 1, Ayu Purwarianti 1, 2, NAIIST, Japan; 2, Institut Teknologi Bandung, Indonesia

This paper presents a novel speech modification method capable of controlling unobservable articulatory parameters based on a statistical feature mapping technique with Gaussian Mixture Models (GMMs). In previous work [1], the GMM-based statistical feature mapping was successfully applied to acoustic-to-articulatory inversion mapping and articulatory-to-acoustic production mapping separately. In this paper, these two mapping frameworks are integrated into a unified framework to develop a novel speech modification system. The proposed system sequentially performs the inversion and the production mapping, making it possible to modify phonemic sounds of an input speech signal by intuitively manipulating articulatory parameters estimated from the input speech signal. We also propose a manipulation method to automatically compensate for unmodified articulatory movements considering inter-dimensional correlation of the articulatory parameters. The proposed system is implemented for a single English speaker and its effectiveness is evaluated experimentally. The experimental results demonstrate that the proposed system is capable of modifying phonemic sounds by manipulating the estimated articulatory movements and higher speech quality is achieved by considering the inter-dimensional correlation in the manipulation.

Speech-Driven Head Motion Synthesis Using Neural Networks

Chuang Ding, Pengcheng Zhu, Lei Xie, Dongmei Jiang, Zhong-Hua Fu; Northwestern Polytechnical University, China

This paper presents a neural network approach for speech-driven head motion synthesis, which can automatically predict a speaker’s head movement from his/her speech. Specifi cally, we realize speech-to-head-motion mapping by learning a multi-layer perceptron from audio-visual broadcast news data. First, we show that a generatively pre-trained neural network signifi cantly outperforms a randomly initialized network and the hidden Markov model (HMM) approach. Second, we demonstrate that the feature combination of log Mel-scale filter-bank (FBank), energy and fundamental frequency (F0) performs best in head motion prediction.
Third, we discover that using long context acoustic information can further improve the performance. Finally, extra unlabeled training data used in the pre-training stage can achieve more performance gain. The proposed speech-driven head motion synthesis approach increases the CCA from 0.299 (the HMM approach) to 0.565 and it can be effectively used in expressive talking avatar animation.

**Text-Independent Voice Conversion Using Speaker Model Alignment Method from Non-Parallel Speech**

Peng Song, Yun Jin, Wenming Zheng, Li Zhao; Southeast University, China

Thu-O-34-3, Poster

In this paper, we propose a novel voice conversion method called speaker model alignment (SMA), which does not require parallel training speech. Firstly, the source and target speaker models, described by Gaussian mixture model (GMM), are trained, respectively. Then, the transformation function of spectral features is learned by aligning the components of source and target speaker models iteratively. Additionally, the transformation function is further combined with GMM, enabling the multiple local mappings, and a local consistent GMM (LCGMM) is also considered for model training to improve the conversion accuracy. Finally, we carry out experiments to evaluate the performance of the proposed method. Objective and subjective experimental results demonstrate that compared with the well-known INCA approach, the proposed method achieves lower spectral distortions and higher correlations, and obtains a significant improvement in perceptual quality and similarity.

**Voice Conversion Using Generative Trained Deep Neural Networks with Multiple Frame Spectral Envelopes**

Ling-Hui Chen, Zhen-Hua Ling, Li-Rong Dai;USTC, China

Wed-P-22-5, Poster

This paper presents a deep neural network (DNN) based spectral envelope conversion method. A global DNN is employed to model the complex non-linear mapping relationship between the spectral envelopes of source and target speakers. The proposed DNN is generatively trained layer-by-layer by cascade of two restricted Boltzmann machines (RBMs) and a bidirectional associative memory (BAM), which are considered as generative models estimated using the contrastive divergence algorithm. Further, multiple spectral envelopes are adopted instead of dynamic features for better modeling using the DNN. The superiority of the proposed method is validated by the subjective experimental results.

**Hierarchical Modeling of F0 Contours for Voice Conversion**

Gerard Sanchez 1, Hanna Silen 1, Jani Nurminen 2, Moncef Gabboury 1, 2; 1Tampere University of Technology, Finland; 2Signalit, Finland

Wed-P-22-6, Poster

Voice conversion systems deal with the conversion of a speech signal to sound as if it was uttered by another speaker. The conversion of the spectral features has attracted a lot of research attention but the conversion of pitch, modeling the speaker-dependent prosody, is often achieved by just controlling the F0 level and range. However, the detailed prosody, including different linguistic units at several distinct temporal scales, can carry a significant amount of speaker identity related information. This paper introduces a new method for the conversion of the prosody, using wavelets to decompose the pitch contour into ten temporal scales ranging from microprosody to the utterance level, which allows modeling the different timings of the prosody phenomena. The prosody conversion is carried out in the wavelet domain, using regression techniques originally developed for the spectral conversion of speech. The performance of the proposed prosody conversion method is evaluated within a real voice conversion system. The results for cross-gender conversion indicate a significant improvement in naturalness when compared to the traditional approach of shifting and scaling the F0 to match the target speaker’s mean and variance.

**Speech Prosody Generation for Text-to-Speech Synthesis Based on Generative Model of F0 Contours**

Kento Kadowaki, Tatsuma Ishihara, Nobukatsu Hojo, Hirokazu Kameoka; University of Tokyo, Japan

Wed-P-22-7, Poster

This paper deals with the problem of generating the fundamental frequency (F0) contour of speech from a text input for text-to-speech synthesis. We have previously introduced a statistical model describing the generating process of speech F0 contours, based on the discrete-time version of the Fujisaki model. One remarkable feature of this model is that it has allowed us to derive an efficient algorithm based on powerful statistical methods for estimating the Fujisaki-model parameters from raw F0 contours. To associate a sequence of the Fujisaki-model parameters with a text input based on statistical learning, this paper proposes extending this model to a context-dependent one. We further propose a parameter training algorithm for the present model based on a decision tree-based context clustering.

**An Iterative Approach to Decision Tree Training for Context Dependent Speech Synthesis**

Xiayu Chen, Yang Zhang, Mark Hasegawa-Johnson; University of Illinois at Urbana-Champaign, USA

Wed-P-22-8, Poster

EDHMM with decision trees is a popular model for parametric speech synthesis. Traditional training procedure constructs the decision trees after observation probability densities have been optimized with the EM algorithm, assuming the state assignment probability does not change much during tree construction. This paper proposes an iterative algorithm that removes the assumption. In the new algorithm, the decision tree construction is incorporated into the EM iteration, with a safeguard procedure that ensures convergence. Evaluation on The Boston University Radio Speech corpus shows that the proposed algorithm can achieve a significantly better optimum in the training set than the original one, and that the advantage is well generalizable to the test set.

**Prosodic Phrasing Modeling for Vietnamese TTS Using Syntactic Information**

Thi Thu Trang Nguyen 1, Albert Rilliard 1, Do Dat Tran 2, Christophe d’Alessandro 1, 2LIMSI, France; 2MICA, Vietnam

Wed-P-22-9, Poster

This research aims at modeling prosodic phrasing for improving the naturalness of Vietnamese (a tonal language) speech synthesis. The proposed phrasing model includes hypotheses on: (i) prosodic structure based on syntactic rules (ii) final lengthening linked to syllabic structures and tone types. Audio files in the analysis corpus are manually transcribed at the syllable level and perceived pauses. Text files are parsed and represented with annotated-
syntax trees. Statistical treatment brings out a correlation between syntactic element boundaries and pause duration. Major breaks may appear at the end of a clause or between predicates or head elements. Other rules between grammatical phrases/words or shorter clauses may trigger minor breaks. Break levels (including ones predicted by syntactic rules) and relative positions of syllables are used to train VTed, an HMM-based Text-To-Speech (TTS) system for Vietnamese. In the synthesis phase, break levels are explicitly inserted while lengthening is applied for last syllables of prosodic phrases. Perceptive testing shows an increase of 0.34 on a 5 point MOS scale, for the new prosodic informed system (3.95/5) compared to the previous TTS system (3.61/5). In the pair-wise comparison test, about 70% of the synthetic voice with the proposed model is preferred to the previous version.

Accent Type and Phrase Boundary Estimation Using Acoustic and Language Models for Automatic Prosodic Labeling
Tomoki Koriyama¹, Hiroshi Suzuki¹, Takashi Nose², Takahiro Shinozaki¹, Takao Kobayashi¹; ¹Tokyo Institute of Technology, Japan; ²Tohoku University, Japan

This paper proposes an automatic prosodic labeling technique for constructing speech database used for speech synthesis. In the corpus-based Japanese speech synthesis, it is essential to use annotated speech data with prosodic information such as phrase boundaries and accent types. However, manual annotation is generally time-consuming and expensive. To overcome this problem, we propose an estimation technique of accent types and phrase boundaries from speech waveform and its transcribed text using both language and acoustic models. We use conditional random field (CRF) for the language model, and HMM for the acoustic model which has shown to be effective in prosody modeling in speech synthesis. By introducing HMM, continuously changing features of F0 contours are modeled well and this results in higher estimation accuracy than conventional techniques that use simple polygon line approximation of F0 contours.

Reconstruction of Mistracked Articulatory Trajectories
Qiang Fang¹, Jianguo Wei², Fang Hu¹; ¹Chinese Academy of Social Sciences, China; ²Tianjin University, China

Kinematic articulatory data are important for researches of speech production, articulatory speech synthesis, robust speech recognition, and speech inversion. Electromagnetic Articulograph (EMA) is a widely used instrument for collecting kinematic articulatory data. However, in EMA experiment, one or more coils attached to articulators are possible to be mistracked due to various reasons. To make full use of the EMA data, we attempt to reconstruct the location of mistracked coils with Gaussian Mixture Model (GMM) regression method. In this paper, we explore how additional information (spectrum, articulatory velocity, etc.) affects the performance of the proposed method. The result indicates that acoustic feature (MFCC) is the most effective additional features that improve the reconstruction performance.

Phone Classification by a Hierarchy of Invariant Representation Layers
Chiyuan Zhang, Stephen Voinea, Georgios Evangelopoulos, Lorenzo Rosasco, Tomaso Poggio; MIT, USA

We propose a multi-layer feature extraction framework for speech, capable of providing invariant representations. A set of templates is generated by sampling the result of applying smooth, identity-preserving transformations (such as vocal tract length and tempo variations) to arbitrarily-selected speech signals. Templates are then stored as the weights of “neurons”. We use a cascade of such computational modules to factor out different types of transformation variability in a hierarchy, and show that it improves phone classification over baseline features. In addition, we describe empirical comparisons of a) different transformations which may be responsible for the variability in speech signals and of b) different ways of assembling template sets for training. The proposed layered system is an effort towards explaining the performance of recent deep learning networks and the principles by which the human auditory cortex might reduce the sample complexity of learning in speech recognition. Our theory and experiments suggest that invariant representations are crucial in learning from complex, real-world data like natural speech. Our model is built on basic computational primitives of cortical neurons, thus making an argument about how representations might be learned in the human auditory cortex.

A Semi-Markov Model for Speech Segmentation with an Utterance-Break Prior
Mark Sinclair, Peter Bell, Alexandra Birch, Fergus McInnes; University of Edinburgh, UK

Speech segmentation is the problem of finding the end points of a speech utterance for passing to an automatic speech recognition (ASR) system. The quality of this segmentation can have a large impact on the accuracy of the ASR system; in this paper we demonstrate that it can have an even larger impact on downstream natural language processing tasks — in this case, machine translation. We develop a novel semi-Markov model which allows the segmentation of audio streams into speech utterances which are optimised for the desired distribution of sentence lengths for the target domain. We compare this with existing state-of-the-art methods and show that it is able to achieve not only improved ASR performance, but also to yield significant benefits to a speech translation task.

Speech Detection in Transient Noises
G. Aneeja, B. Yegnanarayana; IIIT Hyderabad, India

Voice activity detection (VAD) uses a representation of speech derived from spectrum analysis, followed by statistical characterization of speech and degrading noise. Features derived using traditional methods may not be adequate for VAD in the case of transient noises. In this paper, we focus on transient noises where most of the VAD systems in literature do not perform well. A high temporal resolution and high frequency resolution representation
is used to discriminate the transient noises from speech.

The high temporal and frequency resolution representation is achieved by filtering the signal at several single frequencies. The single frequency filtering approach helps to isolate the regions of transient noise in a signal. A time varying threshold is proposed based on the spectral variance and the temporal variance of the speech signal to detect transient noise. The remaining regions are processed by the spectral variance measure for VAD. The results have been compared to the Adaptive Multi-rate (AMR) methods. The performance of proposed method is consistently better due to the instantaneous feature. The percentage of detection of transient noise is higher for the proposed method than the methods reported in the literature.

### Evaluation of Dictionary for Sparse Coding in Speech Processing

Yongjun He¹, Guanglu Sun¹, Guibin Zheng², Jiqing Han²; ¹Harbin Institute of Technology, China; ²Harbin University of Science & Technology, China

![Wed-P-23-4, Poster](image)

As a promising technique, sparse coding has been widely used for the analysis, representation, compression, denoising and separation of speech. To represent a signal accurately and sparsely, a good dictionary which contains elemental signals is urgently desired and many methods have been proposed to obtain such a dictionary. However, there is a lack of reasonable evaluation methods to judge whether a dictionary is good enough. To solve this problem, we define a group of measures for the evaluation of a dictionary. These measures not only address the sparsity and reconstruction error in representation of a signal, but also consider the denoising and separating performances. Experiments show that the proposed measures can make reasonable evaluations.

### Joint Filtering and Factorization for Recovering Latent Structure from Noisy Speech Data

Colin Vaz, Vikram Ramanarayanan, Shrikanth S. Narayanan; University of Southern California, USA

![Wed-P-23-5, Poster](image)

We propose a joint filtering and factorization algorithm to recover latent structure from noisy speech. We incorporate the minimum variance distortionless response (MVDR) formulation within the non-negative matrix factorization (NMF) framework to derive a single, unified cost function for both filtering and factorization. Minimizing this cost function jointly optimizes three quantities—a filter that removes noise, a basis matrix that captures latent structure from noisy speech significantly better than NMF alone or Wiener filtering followed by NMF. Furthermore, PESQ scores show that our algorithm is a viable choice better than NMF alone or Wiener filtering followed by NMF. Furthermore, PESQ scores show that our algorithm is a viable choice.

### A Comparison of Open-Source Segmentation Architectures for Dealing with Imperfect Data from the Media in Speech Synthesis

A. Gallardo-Antolin¹, J.M. Montero², Simon King³; ¹Universidad Carlos III de Madrid, Spain; ²Universidad Politécnica de Madrid, Spain; ³University of Edinburgh, UK

![Wed-P-23-6, Poster](image)

Traditional Text-To-Speech (TTS) systems have been developed using especially-designed non-expressive scripted recordings. In order to develop a new generation of expressive TTS systems in the Simple4All project, real recordings from the media should be used for training new voices with a whole new range of speaking styles. However, for processing this more spontaneous material, the new systems must be able to deal with imperfect data (multi-speaker recordings, background and foreground music and noise), filtering out low-quality audio segments and creating mono-speaker clusters. In this paper we compare several architectures for combining speaker diarization and music and noise detection which improve the precision and overall quality of the segmentation.

### Read and Spontaneous Speech Classification Based on Variance of GMM Supervectors

Taichi Asami, Ryo Masamura, Hirokazu Masataki, Sumitaka Sakauchi; NTT Corporation, Japan

![Wed-P-23-7, Poster](image)

This paper provides a novel method to classify spoken utterances into reading style or spontaneous style. Read/spontaneous speech classification is important for extracting data to train acoustic models for speech recognition from real data in which read speech and spontaneous speech samples are mixed. We analyzed 23,900 reading and 31,988 spontaneous utterances of 30 speakers and found that variance of GMM supervectors in several consecutive utterances can discriminate the reading and spontaneous styles and has less speaker-dependency. Based on this knowledge, our method uses variance of GMM supervectors to classify unknown consecutive utterances into reading style or spontaneous style. Experiments show that our technique can classify 5 consecutive utterances of unknown speakers with over 95% accuracy without any other lexical, phonetic, or prosodic features.

### Co-Channel Speech Detection via Spectral Analysis of Frequency Modulated Sub-Bands

Navid Shokouhi, Seyed Omid Sadjadi, John H.L. Hansen; University of Texas at Dallas, USA

![Wed-P-23-8, Poster](image)

Overlapped-speech is known to degrade performance in automatic speech systems. In this study, a sub-band speech analysis technique is proposed to detect overlapped-speech segments in single-channel multi-speaker scenarios (i.e., co-channel speech). Sub-band signals are obtained by decomposing the input speech using a gammatone filterbank. Filterbank outputs are then used to modulate the frequency argument of a sinusoidal carrier. We show that the spectra of these frequency-modulated signals, namely Gammatone Sub-band Frequency Modulation (GSFM) features, are more disperse in overlapped-speech segments compared to single-speaker regions. We quantify the dispersion rate to obtain a measure for the amount of overlapped speech in a given speech segment. Overlap detection experiments are conducted using the speech separation challenge corpus and GSFM features are compared to commonly used overlap detection features. Detection errors are reduced by a relative 50% across different signal-to-interference values ranging from 0 to 9dB.

### Word-Level Invariant Representations from Acoustic Waveforms

Stephen Voinea, Chiyuan Zhang, Georgios Evangelopoulos, Lorenzo Rosasco, Tomaso Poggio; MIT, USA

![Wed-P-23-9, Poster](image)

Extracting discriminant, transformation-invariant features from raw audio signals remains a serious challenge for speech recog...
tion. The issue of speaker variability is central to this problem, as changes in accent, dialect, gender, and age alter the sound waveform of speech units at multiple levels (phonemes, words, or phrases). Approaches for dealing with this variability have typically focused on analyzing the spectral properties of speech at the level of frames, on par with frame-level acoustic modeling usually applied to speech recognition systems. In this paper, we propose a framework for representing speech at the word level and extracting features from the acoustic, temporal domain, without the need for spectral encoding or preprocessing. Leveraging recent work on unsupervised learning of invariant sensory representations, we extract a signature for a word by first projecting its raw waveform onto a set of templates and their transformations, and then forming empirical estimates of the resulting one-dimensional distributions via histograms. The representation and relevant parameters are evaluated for word classification on a series of datasets with increasing speaker-mismatch difficulty, and the results are compared to those of an MFCC-based representation.

On Closed Form Calculation of Line Spectral Frequencies (LSF)

**Paul Dalsgaard, Ove Andersen; Aalborg University, Denmark**

**Wed P-23-10, Poster**

The mathematical theory of closed form functions for calculating LSFs on the basis of generating functions is presented. Exploiting recurrence relationships in the series expansion of Chebyshev polynomials of the first kind makes it possible to bootstrap iterative LSF-search from a set of characteristic polynomial zeros. The theoretical analysis is based on decomposition into symmetric and anti-symmetric polynomials defined as a series expansion of reduced Chebyshev polynomials of the first kind. Two variants of closed form functions are presented — each characterised by using a recurrence relationship in Chebyshev polynomials. The first exploits the well known three terms recurrence relationships of Chebyshev polynomials. The second hitherto unused recurrence properties of Chebyshev coefficients defining a set of coefficients and zeros used for bootstrapping calculation of LSFs. The theory is tested using bootstrapped calculation of zeros and by evaluating the complexity of the closed form function. The results of the lower complexity calculations show that real axis zeros are within a given iteration tolerance when compared to results of a standard root-finder.

Robust Features for Content-Based Audio Copy Detection

**Chahid Ouali**, Pierre Dumouchel; 1École de Technologie Supérieure, Canada; 2CRIM, Canada

**Wed P-23-11, Poster**

In this paper, we present the latest improvements on spectrogram-matrix based fingerprinting system for detecting transformed audio copies. In particular, we experiment with two feature parameters derived using global and local spectrogram averages and show that combining results from these two feature parameters significantly improves performance. We test our system on TRECVID 2010 content-based copy detection dataset. Experimental results show the robustness of our method against various audio distortions. The proposed method reduces the minimal Normalized Detection Cost Rate (min NDCR) by 23% and improves localization accuracy by 24% compared to a state-of-the-art audio fingerprinting system. Our system achieves the lowest min NDCR for MP3 compression transformation with a relative improvement of 60%.

### Binaural Deep Neural Network Classification for Reverberant Speech Segregation

**Yi Jiang**, DeLiang Wang; RunSheng Liu; Tsinghua University, China; Ohio State University, USA

**Wed-P-23-12, Poster**

While human listening is robust in complex auditory scenes, current speech segregation algorithms do not perform well in noisy and reverberant environments. This paper addresses the robustness in binaural speech segregation by employing binary classification based on deep neural networks (DNNs). We systematically examine DNN generalization to untrained configurations. Evaluations and comparisons show that DNN based binaural classification produces superior segregation performance in a variety of multisource and reverberant conditions.

### Poster Session 24: Feature Extraction and Modeling for ASR

**Max Atria Gallery**

**16:00 – 18:00, Wednesday 17 September 2014**

**Chair: Jinyu Li**

### Investigating NMF Speech Enhancement for Neural Network Based Acoustic Models

**Jürgen T. Geiger**, Jort F. Gemmeke; Björn Schuller; Gerhard Rigoll; Technische Universität München, Germany; Katholieke Universiteit Leuven, Belgium; Imperial College London, UK

**Wed P-24-1, Poster**

In the light of the improvements that were made in the last years with neural network-based acoustic models, it is an interesting question whether these models are also suited for noise-robust recognition. This has not yet been fully explored, although first experiments confirm this question. Furthermore, preprocessing techniques that improve the robustness should be re-evaluated with these new models. In this work, we present experimental results to address these questions. Acoustic models based on Gaussian mixture models (GMMs), deep neural networks (DNNs), and long short-term memory (LSTM) recurrent neural networks (which have an improved ability to exploit context) are evaluated for their robustness after clean or multi-condition training. In addition, the influence of non-negative matrix factorization (NMF) for speech enhancement is investigated. Experiments are performed with the Aurora-4 database and the results show that DNNs perform slightly better than LSTMs and, as expected, both beat GMMs. Furthermore, speech enhancement is capable of improving the DNN result.

### Automatic Speech Feature Classification for Children with Cochlear Implants

**Jason Lilley**, James Mahshie; H. Timothy Bunnell; Nemours Biomedical Research, USA; George Washington University, USA

**Wed P-24-2, Poster**

Following pediatric cochlear implantation, it is important to monitor the recipient’s auditory perception skills to (a) determine appropriate therapeutic strategies, and (b) guide adjustment (mapping) of how the CI stimulation is distributed over its channels. Approaches for dealing with this variability are usually applied to speech recognition systems. In this paper, we present a framework for representing speech at the word level and extracting features from the acoustic, temporal domain, without the need for spectral encoding or preprocessing. Leveraging recent work on unsupervised learning of invariant sensory representations, we extract a signature for a word by first projecting its raw waveform onto a set of templates and their transformations, and then forming empirical estimates of the resulting one-dimensional distributions via histograms. The representation and relevant parameters are evaluated for word classification on a series of datasets with increasing speaker-mismatch difficulty, and the results are compared to those of an MFCC-based representation.
Sequential Maximum Mutual Information Linear Discriminant Analysis for Speech Recognition
Yuuki Tachioka¹, Shintaro Watanabe², Jonathan Le Roux², John R. Hershey²; ¹Mitsubishi Electric, Japan; ²MERL, USA

Linear discriminant analysis (LDA) is a simple and effective feature transformation technique that aims to improve discriminability by maximizing the ratio of the between-class variance to the within-class variance. However, LDA does not explicitly consider the sequential discriminative criterion which consists in directly reducing the errors of a speech recognizer. This paper proposes a simple extension of LDA that is called sequential LDA (sLDA) based on a sequential discriminative criterion computed from the Gaussian statistics, which are obtained from sequential maximum mutual information (MMI) training. Although the objective function of the proposed LDA can be regarded as a specific case of various discriminative feature transformation techniques (for example, f-MPE or the bottom layer of a neural network), the transformation matrix can be obtained as the closed-form solution to a generalized eigenvalue problem, in contrast to the gradient-descent-based optimization methods usually used in these techniques. Experiments on LVCSR (Corpus of Spontaneous Japanese) and noisy speech recognition task (2nd CHiME challenge) show consistent improvements from standard LDA due to the sequential discriminative training. In addition, the proposed method, despite its simple and fast computation, improved the performance in combination with discriminative feature transformation (f-bMMI), perhaps by providing a good initialization to f-bMMI.

Model and Feature Based Compensation for Whispered Speech Recognition
Shabnam Ghaffarzadegan, Hynek Bořil, John H.L. Hansen; University of Texas at Dallas, USA

This study proposes model and feature based strategies for automatic whispered speech recognition. Our goal is to compensate for the mismatch between neutral-trained recognizer models and parameters of whispered speech. We propose a pseudo-whisper generation from neutral speech samples for efficient acoustic model adaptation. The scheme is based on the popular Vector Taylor Series (VTS) algorithm. In the first step, a ‘background’ model capturing a rough estimate of the target whispered speech characteristics from a small amount of whispered data is trained. Second, the target background model is utilized in the VTS strategy to establish broad phone classes (consonants and vowels) transformations for individual neutral utterances and transform them towards whisper. Finally, these pseudo-whisper samples are used to adapt neutral recognizer models towards whisper. This approach is evaluated together with Vocal Tract Normalization (VTN) and Shift frequency transforms and show to greatly benefit recognition performance compared to a traditional whisper-adaptation approach. The absolute WER on the closed speakers whisper scenario has been reduced from 17.3% to 8.4% and the open speakers scenario from 27.7% to 17.5%.

Post-Masking: A Hybrid Approach to Array Processing for Speech Recognition
Amir R. Moghi, Bikhsha Raj, Richard M. Stern; Carnegie Mellon University, USA

In the context of array processing for speech and audio applications, linear beamforming has long been the approach of choice, for reasons including good performance, robustness and analytical simplicity. Nevertheless, various nonlinear techniques, typically based on the study of auditory scene analysis, have also been of interest. The class of techniques known as time-frequency (T-F) masking, in particular, shows promise; T-F masking is based on accepting or rejecting individual time-frequency cells based on some estimate of local signal quality. While these approaches have been shown to outperform linear beamforming in two-sensor arrays, extensions to larger arrays have been few and unsuccessful. This paper seeks to gain a deeper understanding of the limitations of T-F masking in larger arrays and to develop an approach to overcome them. It is shown that combining beamforming and masking can bring the benefits of masking to larger arrays. As a result, a hybrid beamforming masking approach, called post-masking, is developed that improves upon the performance of MMSE beamforming (and can be used with any beamforming technique). Post-masking extends the benefits of masking up to arrays of six elements or more, with the potential for even greater improvement in the future.

ASR Feature Extraction with Morphologically-Filtered Power-Normalized Cochleograms
F. de-la-Calle-Silos¹, F.J. Valverde-Albacete², A. Gallardo-Antolin¹, C. Peláez-Moreno¹; ¹Universidad Carlos III de Madrid, Spain; ²Universidad Nacional de Educación a Distancia, Spain

In this paper we present advances in the modeling of the masking behavior of the Human Auditory System to enhance the robustness of the feature extraction stage in Automatic Speech Recognition. The solution adopted is based on a non-linear filtering of a spectro-temporal representation applied simultaneously on both the frequency and time domains, by processing it using mathematical morphology operations as if it were an image. A particularly important component of this architecture is the so-called structuring element: biologically-based considerations are addressed in the present contribution to design an element that closely resembles the masking phenomena taking place in the cochlea. The second feature of this contribution is the choice of underlying spectro-temporal representation. The best results were achieved by the representation introduced as part of the Power Normalized Cepstral Coefficients together with a spectral subtraction step. On the Aurora 2 noisy continuous digits task, we report relative error reductions of 18.7% compared to PNCC and 39.5% compared to MFCC.
Should Deep Neural Nets Have Ears? The Role of Auditory Features in Deep Learning Approaches

Angel Mario Castro Martínez, Niko Moritz, Bernd T. Meyer; Carl von Ossietzky Universität Oldenburg, Germany

Wed-P-24-7, Poster

Features inspired by the auditory system have previously demonstrated improvement in automatic speech recognition (ASR). Similarly, the use of Deep Neural Networks (DNN) was found to outperform classic approaches to ASR in many conditions. Since DNNs have the potential to learn the task-relevant features from a conventional filter bank output, we investigate if the combination of auditory features and deep learning should be preferred over self-learned patterns. Specifically, noise-robust Gabor features and Amplitude Modulation Filter-Bank (AMFB) features, highly invariant against reverberation, are used as input to a state-of-the-art ASR system incorporating DNN processing. On the Aurora-4 task, both mel-frequency cepstral coefficients (MFCC) and filter bank (FBank) features are outperformed in many acoustic conditions through auditory processing, yielding average relative improvements of up to 6% over MFCC and 21% over the commonly used DNN-FBank setup. This highlights the mutual benefit of auditory signal processing and recent advances in machine learning.

Extending Limabeam with Discrimination and Coarse Gradients

Charles Fox, Thomas Hain; University of Sheffield, UK

Wed-P-24-8, Poster

Limabeam is an approach to multi-microphone array processing for ASR which makes minimal assumptions about system geometry, instead searching for filters to maximise output likelihoods under a speech model. The first results of Limabeam on the AMI meeting corpus are given, then two extensions of the algorithm for this corpus. First, it is shown that the original local gradient following sticks in local minima, and a coarser gradient is used. Second, a new discriminative objective function is provided to handle mismatched silence models. The extensions are based on examination of 2D receptive fields and 2D likelihood maps which are novel near-field analogs of radial beamformer response patterns, but do not show radial symmetry and have many local minima. The extended Limabeam improves WER on TDOA baselines on the AMI corpus, by 1% rel. when both are adapted with 15% rel. when both adapted with ground-truth.

Generation of F0 Contour Using Deep Boltzmann Machine and Twin Gaussian Process Hybrid Model for Bengali Language

Sankar Mukherjee, Shyamal Kumar Das Mandal; IIT Kharagpur, India

Wed-P-24-9, Poster

In Text to Speech synthesis system F0 contour plays an important role in conveying prosodic information but the process of synthesizing F0 contours from the underlying linguistic information using deep architecture has not been investigated in case of Bengali languages. This paper describes a method for synthesizing F0 contours of Bengali readout speech from the textual features of input text using Deep Boltzmann Machine (DBM) and Twin Gaussian Process (TGP) hybrid model. DBM will capture the high-level linguistic structure of input text and improve the prediction accuracy when plugged into the TGP model. Unlike Gaussian Process (GP) models which only focus on the prediction of a single output (F0), TGP can generalize across multiple outputs (F0, delta-delta F0) by encoding relations between both inputs and outputs with GP priors. The performance of the proposed method is evaluated and compared with other available methods using objective and perceptual listening tests and the results are found to be satisfactory.

Room Localization for Distant Speech Recognition

Juan A. Morales-Cordovilla, Hannes Pessentheiner, Martin Hagmüller, Gernot Kubin; Technische Universität Graz, Austria

Wed-P-24-10, Poster

The problem of room localization is to determine where, in a multi-room environment, a person is producing a speech utterance. In our work, we are exploiting the information gained from a network of microphones installed all over a house, where the lack of calibration of the microphone energies creates an additional challenge. This paper compares room localizers based on different features (such as energy and cross-correlation between microphones) and classifiers (such as neural networks and discriminative analysis). In order to evaluate the different room localizers in terms of word accuracy this paper also presents a complete distant speech recognition system which tries to take advantage of synergy between the different components without using any oracle information. Finally, the system is analyzed in terms of computational and time resources.

Posterior-Based Sparse Representation for Automatic Speech Recognition

Sara Bahadini, Afsaneh Asaei, David Imseng, Hervé Bourlard; Idiap Research Institute, Switzerland

Wed-P-24-11, Poster

Posterior features have been shown to yield very good performance in multiple contexts including speech recognition, spoken term detection, and template matching. These days, posterior features are usually estimated at the output of a neural network. More recently, sparse representation has also been shown to potentially provide additional advantages to improve discrimination and robustness. One possible instance of this, is referred to as exemplar-based sparse representation. The present work investigates how to exploit sparse modelling together with posterior space properties to further improve speech recognition features. In that context, we leverage exemplar-based sparse representation, and propose a novel approach to project phone posterior features into a new, high-dimensional, sparse feature space. In fact, exploiting the properties of posterior spaces, we generate, new, high-dimensional, linguistically inspired (sub-)phone and words), posterior distributions. Validation experiments are performed on the Phonebook (isolated words) and HIWIRE (continuous speech) databases, which support the effectiveness of the proposed approach for speech recognition tasks.

NOTES
Keynote 5: Li Deng
Garnet 213-218
08:30 – 09:30, Thursday 18 September 2014
Chair: Douglas O’Shaughnessy

Achievements and Challenges of Deep Learning — From Speech Analysis and Recognition to Language and Multimodal Processing

Li Deng; Microsoft, USA

Keynote 5

Artificial neural networks have been around for over half a century and their applications to speech processing have been almost as long, yet it was not until year 2010 that their real impact had been made by a deep form of such networks, built upon part of the earlier work on (shallow) neural nets and (deep) graphical models developed by both speech and machine learning communities. This keynote will first reflect on the path to this transformative success, sparked by speech analysis using deep learning methods on spectrogram-like raw features and then progressing rapidly to speech recognition with increasingly larger vocabularies and scale. The role of well-timed academic-industrial collaboration will be highlighted, so will be the advances of big data, big compute, and the seamless integration between the application-domain knowledge of speech and general principles of deep learning. Then, an overview will be given on sweeping achievements of deep learning in speech recognition since its initial success in 2010 (as well as in image recognition and computer vision since 2012). Such achievements have resulted in across-the-board, industry-wide deployment of deep learning. The final part of the talk will look ahead towards stimulating new challenges of deep learning — making intelligent machines capable of not only hearing (speech) and seeing (vision), but also of thinking with a "mind"; i.e., reasoning and inference over complex, hierarchical relationships and knowledge sources that comprise a vast number of entities and semantic concepts in the real world based in part on multisensory data from the user. To this end, language and multimodal processing — joint exploitation and learning from text, speech/audio, and image/video — is evolving into a new frontier of deep learning, beginning to be embraced by a mixture of research communities including speech and spoken language processing, natural language processing, computer vision, machine learning, information retrieval, cognitive science, artificial intelligence, and data/knowledge management. A review of recent published studies will be provided on deep learning applied to selected language and multimodal processing tasks, with a trace back to the relevant early connectionist modeling and neural network literature and with future directions in this new exciting deep learning frontier discussed and analyzed.

Oral Session 33: Spoken Term Detection for Low-Resource Languages I
Garnet 213-218
10:00 – 12:00, Thursday 18 September 2014
Chairs: Hermann Ney and Brian Kingsbury

Query-by-Example Spoken Term Detection on Multilingual Unconstrained Speech

Xavier Anguera 1, Luis Javier Rodriguez-Fuentes 2, Igor Szöke 3, Andi Buzo 4, Florian Metze 5, Mikel Penagarikano 2; 1Telefónica I+D, Spain; 2Universidad del País Vasco, Spain; 3Brno University of Technology, Czech Republic; 4Universitatea Politehnica din București, Romania; 5Carnegie Mellon University, USA

Thu-O-33-1

As part of the MediaEval 2013 benchmark evaluation campaign, the objective of the Spoken Web Search (SWS) task was to perform Query-by-Example Spoken Term Detection (QbE-STD) using audio queries in a low-resource setting. After two successful editions and a continuously growing interest in the scientific community, a special effort was made in SWS 2013 to prepare a challenging database, including speech in 9 different languages with diverse environment and channel conditions. In this paper, first we describe the database and the performance metrics. Then, we briefly review the algorithmic approaches followed by participants and present and discuss the obtained performances, which demonstrate the feasibility of the proposed task, even under such challenging conditions (multiple languages and unconstrained acoustic conditions). Finally, we analyze the fusion of the top-performing systems, which achieved a 30% relative improvement over the best single system in the evaluation, proving that a variety of approaches can be effectively combined to bring complementary information in the search for queries.

A Comparison of Multiple Methods for Rescoring Keyword Search Lists for Low Resource Languages

Victor Soto 1, Lidia Mangu 2, Andrew Rosenberg 3, Julia Hirschberg 1; 1Columbia University, USA; 2IBM T.J. Watson Research Center, USA; 3CUNY Graduate Center, USA

Thu-O-33-2

We review the performance of a new two-stage cascaded machine learning approach for rescoring keyword search output for low resource languages. In the first stage Confusion Networks (CNs) are rescoring for improved Automatic Speech Recognition (ASR) by reranking the arcs of each confusion bin. In the second stage we generate keyword search hypotheses from the rescored ASR output and rescore them using logistic regression classifiers to detect true hits and false alarms. We compare the performance of our system with state of the art rescoring techniques, including probability of false alarm normalization, exponential normalization, rank-normalized posterior scores and sum-to-one normalization and show promising results. Experimental validation is performed using the Term Weighted Value (TWV) metric on four corpora from the IARPA-Babel program for keyword search on low resource languages, including Assamese, Bengali, Lao and Zulu.

NOTES
Subword and Phonetic Search for Detecting Out-of-Vocabulary Keywords

Damianos Karakos, Richard Schwartz; Raytheon BBN Technologies, USA

Thu-O-33-3

We compare several approaches, separately and together, for spotting of out-of-vocabulary (OOV) keywords, in terms of their ATWV scores. We considered three types of recognition units (whole words, syllables, and subwords of different lengths) and two basic search strategies (whole-unit, fuzzy phonetic search). In all cases, the search was performed by collapsing the recognition lattice into a consensus network, either in terms of the recognized whole units, or by first splitting the recognized units into phonemes. We ran experiments on five languages, for which the language model and vocabulary were derived from only 10 hours of transcriptions (70k-100k words of text), resulting in keyword OOV rates varying from 10% to 63% on new data, depending on the language. Our conclusions were that: 1) In all cases, the fuzzy phonetic search on phoneme-split lattices is better than searching for the whole units, 2) The syllable units are the best of the subword units for OOV keyword detection using fuzzy phonetic search, and 3) These methods combine very well, sometimes resulting in ATWV scores for OOV terms which are not too far below those of IV terms.

An In-Depth Comparison of Keyword Specific Thresholding and Sum-to-One Score Normalization

Yun Wang, Florian Metze; Carnegie Mellon University, USA

Thu-O-33-4

The quality of a spoken term detection (STD) system critically depends on the choice of a “thresholding” function, which is used to determine whether to output a candidate detection or not based on its score. In the context of the IARPA Babel program and the NIST OpenKWS evaluation series, the penalty for missing an occurrence depends on the frequency of the keyword, so it is desirable either to apply different thresholds to different keywords, or to normalize the scores before applying a global threshold. This paper compares two widely used thresholding algorithms: keyword specific thresholding (KST) and sum-to-one score normalization (STO), analyzes the difference in their performance in detail, and recommends the use of the “estimated KST” algorithm.

Graph-Based Re-Ranking Using Acoustic Feature Similarity Between Search Results for Spoken Term Detection on Low-Resource Languages

Hung-yi Lee, Yu Zhang, Ekapol Chuangsuwanich, James R. Glass; MIT, USA

Thu-O-33-5

Acoustic feature similarity between search results has been shown to be very helpful for the task of spoken term detection (STD). A graph-based re-ranking approach for STD has been proposed based on the concept that search results, which are acoustically similar to other results with higher confidence scores, should have higher scores themselves. In this approach, the similarity between all search results of a given term are considered as a graph, and the confidence scores of the search results propagate through this graph. Since this approach can improve STD results without any additional labelled data, it is especially suitable for STD on languages with limited amounts of annotated data. However, its performance has not been widely studied on benchmark corpora. In this paper, we investigate the effectiveness of the graph-based re-ranking approach on limited language data from the IARPA Babel program. Experiments on the low-resource languages, Assamese, Bengali and Lao, show that graph-based re-ranking improves STD systems using fuzzy matching, and lattices based on different kinds of units including words, subwords, and hybrids.

Developing STT and KWS Systems Using Limited Language Resources

Viet-Bac Le 1, Lori Lamel 2, Abdel Messaoudi 1, William Hartmann 2, Jean-Luc Gauvain 2, Cécile Woehrling 1, Julien Despres 1, Anindya Roy 2; 1Vocapia Research, France; 2LIMSI, France

Thu-O-33-6

This paper presents recent progress in developing speech-to-text (STT) and keyword spotting (KWS) systems for the 2014 IARPA-Babel evaluation. Systems have been developed for the limited language pack condition for four of the five development languages in this program phase: Assamese, Bengali, Haitian Creole and Zulu. The systems have several novel characteristics that support rapid development of KWS systems. On the STT side different acoustic units are explored based on phonemic or graphemic representations, and system combination is used to improve STT performance. The acoustic models are trained on only 10 hours of speech data with manual transcriptions, completed with unsupervised training on additional untranscribed data. Both word and subword units (morphologically decomposed, syllables, phonemes) are used for KWS. The KWS systems are based on the multi-hypotheses produced by a consensus network decoding or searching word lattices. The word error rates of the individual STT systems are on the order of 50–60%, and the KWS systems obtain Maximum Term Weighted Values ranging from 30–45% for all keywords (in-vocabulary and out-of-vocabulary (OOV)). Sub-word units are shown to be successful at locating some of the OOV keywords, and system combination improves system performance.

Oral Session 34: Voice Conversion

Period 202-203
10:00 – 12:00, Thursday 18 September 2014
Chairs: Junichi Yamagishi and Olivier Rosec

GMM-Based Bandwidth Extension Using Sub-Band Basis Spectrum Model

Yamato Ohtani, Masatsune Tamura, Masahiro Morita, Masami Akamine; Toshiba, Japan

Thu-O-34-1

This paper describes a novel GMM-based bandwidth extension (BWE) method based on a sub-band basis spectrum model (SBM), in which each dimensional component represents a specific acoustic space in the frequency domain. The proposed method can achieve the BWE from a speech data with an arbitrary frequency bandwidth while the conventional methods perform the conversion from a fixed narrowband data. In the proposed method, we train a GMM with SBM parameters extracted from wideband spectra in advance. An input signal with a limited frequency band is converted into a wideband signal by estimating high-band SBM components from low-band SBM components of the input signal based on the GMM. The results of some objective and subjective evaluations show that the proposed method extends bandwidth of speech data robustly.
A Mel-Cepstral Analysis Technique Restoring High Frequency Components from Low-Sampling-Rate Speech
Kazushiro Nakamura, Kei Hashimoto, Keichiro Oura, Yoshihiko Nankaku, Keiichi Tokuda; Nagoya Institute of Technology, Japan
Thu-O-34-2
In statistical speech synthesis, the quality of the synthesized speech depends on the quality of training data. As the sampling rate of speech is one of the effective factors, speech data has been recently recorded at a high sampling rate. However, the sampling rates of speech data recorded in the past or collected from the internet were often low. Therefore, to use these speech data effectively for model training, we propose a mel-cepstral analysis technique that restores missing high frequency components from low-sampling-rate speech with a statistical approach. In this technique, high-sampling-rate speech waveforms are modeled directly by integrating feature extraction and modeling processes. This framework makes it possible to optimize whole processes on the basis of an integrated objective function. Then, mel-cepstral coefficients are estimated from the low-sampling-rate speech by using the model as a prior distribution. Experimental results show that the proposed method improved the quality of synthesized speech.

A Comparative Study of Spectral Transformation Techniques for Singing Voice Synthesis
S.W. Lee 1, Zhizheng Wu 2, Miningh Dong 1, Xiaohai Tian 2, Haizhou Li 1; 1A*STAR, Singapore; 2Nanyang Technological University, Singapore
Wed-F-32-27
Studies show that professional singing matches well the associated melody and typically exhibits spectra different from speech in resonance tuning and singing formant. Therefore, one of the important topics in speech-to-singing conversion is to characterize the spectral transformation between speech and singing. This paper extends two types of spectral transformation techniques, namely voice conversion and model adaptation, and examines their performance. For the first time, we carry out a comparative study over four singing voice synthesis techniques. The experiments on various data sizes reveal that maximum-likelihood Gaussian mixture model (ML-GMM) of voice conversion always delivers the best performance in terms of spectral estimation accuracy; while model adaptation generates the best singing quality in all cases. When a large dataset is available, both techniques achieve the highest similarity to target singing. With a small dataset, the highest similarity is obtained by ML-GMM. It is also found that the music context-dependent modeling in adaptation, in which detailed partition of transform space is involved, leads to pleasant singing spectra.

Application of Matrix Variate Gaussian Mixture Model to Statistical Voice Conversion
Daisuke Saito, Hidenobu Doi, Nobuaki Minematsu, Keiichi Hirose; University of Tokyo, Japan
Thu-O-34-4
This paper describes a novel approach to construct a mapping function between a given speaker pair using probability density functions (PDF) of matrix variate. In voice conversion studies, two important functions should be realized: 1) precise modeling of both the source and target feature spaces, and 2) construction of a proper transform function between these spaces. Voice conversion based on Gaussian mixture model (GMM) is the de facto standard because of their flexibility and easiness in handling. In GMM-based approaches, a joint vector space of the source and target is first constructed, and the joint PDF of the two vectors is modeled as GMM in the joint vector space. The joint vector approach mainly focuses on precise modeling of the ‘joint’ feature space, and does not always construct a proper transform between two feature spaces. In contrast, the proposed method constructs the joint PDF as GMM in a matrix variate space whose row and column respectively correspond to the two functions, and it has potential to precisely model both the characteristics of the feature spaces and the relation between the source and target spaces.

Joint Nonnegative Matrix Factorization for Exemplar-Based Voice Conversion
Zhizheng Wu, Eng Song Ching, Haizhou Li; Nanyang Technological University, Singapore
Thu-O-34-5
Recently, exemplar-based sparse representation methods have been proposed for voice conversion. These methods reconstruct a target spectrum through a weighted linear combination from a set of basis spectra, called exemplars. To include temporal constraint, multiple-frame exemplars are employed when estimating the linear combination weights, namely activations, by the nonnegative matrix factorization technique with a sparsity constraint. In practice, low-resolution mel-scale filter bank energies rather than high-resolution spectra are employed to estimate the activations in order to reduce computational cost and memory usages. However, the conversion performance degrades due to the loss of the spectral details in the low-resolution representations. In this study, we propose a joint nonnegative matrix factorization technique to estimate the activations using both the low- and high-resolution features simultaneously. In this way, we include temporal information by using multiple-frame low-resolution exemplars for computational efficiency and one-frame high-resolution exemplars to improve spectral details at the same time. The VOICES database was employed to assess the performance of the proposed method. The experiments confirmed the effectiveness of the proposed method over conventional nonnegative matrix factorization method in term of both objective spectral distortion and subjective evaluation.

Statistical Singing Voice Conversion with Direct Waveform Modification based on the Spectrum Differential
Kazushiro Kobayashi, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura; NAIST, Japan
Thu-O-34-6
This paper presents a novel statistical singing voice conversion (SVC) technique with direct waveform modification based on the spectrum differential that can convert voice timbre of a source singer into that of a target singer without using a vocoder to generate converted singing voice waveforms. SVC makes it possible to convert singing voice characteristics of an arbitrary source singer into those of an arbitrary target singer. However, speech quality of the converted singing voice is significantly degraded compared to that of a natural singing voice due to various factors, such as analysis and modeling errors in the vocoder-based framework. To alleviate this degradation, we propose a statistical conversion process that directly modifies the signal in the waveform domain by estimating the difference in the spectra of the source and target singers’ singing voices. The differential spectral feature is directly estimated using a differential Gaussian mixture model (GMM) that is analytically derived from the traditional GMM used as a conversion model in the conventional SVC. The experimental results demonstrate that the proposed method makes it possible

NOTES
to significantly improve speech quality in the converted singing voice while preserving the conversion accuracy of singer identity compared to the conventional SVC.

Oral Session 35: Speech and Audio Segmentation and Classification
Peridot 204-205
10:00 – 12:00, Thursday 18 September 2014
Chairs: Lin-shan Lee and Michael Picheny

Detecting Proximity from Personal Audio Recordings
Daniel P.W. Ellis¹, Hironyaki Satoh², Zhuo Chen¹; ¹Columbia University, USA; ²University of Tokyo, Japan
Thu-O-35-1

It is common to be carrying an advanced computational device with a microphone — a smartphone — on your person at virtually all times. One application this makes possible is to automatically detect when individuals are in close proximity by detecting the similarity between the acoustic ambience recorded by body-worn mics. This paper investigates two techniques for proximity detection on a database of personal audio recordings made by six participants in a poster presentation session. We show that cross-correlation between 10 s windows is effective for detecting when individuals are close enough to be in conversation, and that using a fingerprinting approach based on acoustic landmarks is comparably accurate for this task, while at the same time being much more efficient, privacy-preserving, and viable for detecting proximity between a large number of body-worn devices.

Acoustic Event Detection and Localization with Regression Forests
Huy Phan, Marco Maaß, Radoslaw Mazur, Alfred Mertins; Universität zu Lübeck, Germany
Thu-O-35-2

This paper proposes an approach for the efficient automatic joint detection and localization of single-channel acoustic events using random forest regression. The audio signals are decomposed into multiple densely overlapping superframes annotated with event class labels and their displacements to the temporal starting and ending points of the events. Using the displacement information, a multivariate random forest regression model is learned for each event category to map each superframe to continuous estimates of onset and offset locations of the events. In addition, two classifiers are trained using random forest classification to classify superframes of background and different event categories. On testing, based on the detection of category-specific superframes using the classifiers, the learned regressor provides the estimates of onset and offset locations in time of the corresponding event. While posing event detection and localization as a regression problem is novel, the quantitative evaluation on ITC-First database of highly variable acoustic events shows the efficiency and potential of the proposed approach.

Multi-Source Posteriors for Speech Activity Detection on Public Talks
Marc Ferràs, Hervé Bourlard; Idiap Research Institute, Switzerland
Thu-O-35-3

Speech activity detection (SAD) is a conceptually simple task that still poses serious challenges for speech processing in a large variety of scenarios. Current energy-based and model-based approaches tend to directly segment speech and non-speech classes, but are not robust enough to non-stationary noise. In this paper, we use a multi-source activity detection (MSAD) approach to SAD by finding the activity levels of speech and a set of non-speech acoustic sources. Public talks such as TED involve a large variety of non-speech audio that is difficult to handle with standard SAD systems. We evaluate the effect of using either the proposed MSAD system versus a tailored version of the popular SHOUT SAD systems. We evaluate the effect of using either the proposed MSAD system versus a tailored version of the popular SHOUT SAD systems. We evaluate the effect of using either the proposed MSAD system versus a tailored version of the popular SHOUT SAD systems. We evaluate the effect of using either the proposed MSAD system versus a tailored version of the popular SHOUT SAD systems.

Notes
comprehensive statistical analysis, showing that the suggested classifier generalizes, and revealing the role and the statistical importance of the different spoken tasks and their respective vocal features.

We demonstrate classification accuracy of about 80% for CTRL vs. MCI and MCI vs. AD, and 87% for CTRL vs. AD, all shown to generalize. This provides an evidence for potential use for automatic detection of very early dementia.

On the Acoustic Environment of a Neonatal Intensive Care Unit: Initial Description, and Detection of Equipment Alarms

Ganna Raboshchuk1, Climent Nadeu1, Omid Ghabahi1, Sergi Solve1, Blanca Muñoz Mahamud2, Ana Riverola de Veciana2, Santiago Navarro Hervas2; 1Universitat Politècnica de Catalunya, Spain; 2Hospital Sant Joan de Déu Barcelona, Spain
Thu-O-355

The acoustic environment of a typical neonatal intensive care unit (NICU) is very rich and may contain a large number of different sounds, which come either from the equipment or from the human activities taking place in it. There exists a medical concern about the effect of that acoustical environment on preterm infants, since loud sounds or particular sounds may be harmful for their further neurological development. In this work, first of all, an initial description of the acoustic characteristics of the NICU has been carried out using a set of diverse recordings produced with microphones placed both inside and outside an incubator. Then, the work has focused on detection of the most relevant types of sounds. In this paper, after describing the recorded database and the acoustic environment, preliminary experiments for detection of the acoustic alarms of devices are reported. The proposed detection system is based on Deep Belief Networks (DBN). The experimental results show that the DBN-based system is able to achieve better results than a baseline GMM-based system.

Oral Session 36: Language Acquisition

Peridot 201
10:00 – 12:00, Thursday 18 September 2014
Chairs: Louis ten Bosch and Mark Huckvale

Non-Native Perception of Regionally Accented Speech in a Multitalker Context

Robert Allen Fox, Ewa Jacewicz, Florence Hardjono; Ohio State University, USA
Thu-O-36-1

Noisy listening conditions are challenging to non-native listeners who typically perform poorly while attending to several competing talkers. This study examined whether non-native listeners are able to utilize dialect-related cues in the target and in the masking speech, even if they do not reach the proficiency level of the native listeners. 35 Indonesian-English bilinguals residing in the United States were presented with speech stimuli from two American English dialects, General American English and Southern American English, which were systematically varied both in the target sentences and in 2-talker masking babble at three sound-to-noise ratios (SNR). We found that the non-native listeners were (1) sensitive to dialect-specific phonetic details in speech of competing talkers and (2) performed in a manner similar to native listeners despite their apparent deficit. However, their performance differed significantly when the speech levels of the competing talkers were equal (0 dB SNR). The differential sensitivity of non-native listeners may reflect their inability to separate utterances of competing talkers when there is not enough contrast in their voice levels. In turn, the lack of sufficient contrast may reduce their ability to benefit from the phonetic-acoustic details necessary to encode the signal and comprehend a message.

A Crosslinguistic and Acquisitional Perspective on Intonational Rises in French

Giuseppina Turco1, Elisabeth Delais-Roussarie2; 1Université Konstanz, Germany; 2LFL (UMR 7110), France
Thu-O-36-2

This study compares rising contours produced in the context of contrastive topics by French natives and by low and high proficient learners of French with German as mother tongue. Results show a systematic pattern for French natives who mostly produced a final rise LH, and hardly ever a bridge accent on the whole phrase. Our results on French natives seem to support earlier claims that tonal patterns with late dip alignments may be recruited for encoding contrast meaning. Results on French learners show a development in the acquisition of the prosody-semantics mapping principles (shifting the accent position from the phrase-initial mon to the phrase-final image) and, not surprisingly, differences in the phonetic implementation of the final rises. Crucially, the impact of phonological and phonetic transfer is more complex than expected: text-to-tune associations are not easy to re-programme when a new accent location has to be learnt. However, once the phonology is learnt, the phonetic implementation starts being problematic.

Error Patterns of Mandarin Disyllabic Tones by Japanese Learners

Jung-Yueh Tu, Yuwen Hsung, Min-Da Wu, Yao-Ting Sung; National Taiwan Normal University, Taiwan
Thu-O-36-3

Previous studies on Mandarin tone production indicate that there is no agreement on which tones are most difficult for L2 learners. Much of previous research on L2 learning of Mandarin tones has focused on monosyllables. In modern Mandarin, however, it is disyllabic words that dominate the vocabulary. This research investigates the production of Mandarin disyllabic tones by Japanese learners. In the current study, 25 Japanese learners of Mandarin were requested to produce 80 Mandarin disyllabic words with all tonal combinations (except for the neutral tone). The overall results showed a hierarchy of difficulty: Tone 3 > Tone 2 > Tone 1 = Tone 4. Most errors in the first syllable were found for Tone 2 and Tone 3 when followed by Tone 1 or Tone 4 (both start with a high pitch). In the second syllable, most errors were found for Tone 3 (misproduced as Tone 2). The findings are discussed in terms of the phonetic nature of Mandarin lexical tones and the interference from Japanese phonology.

Infant-Directed Speech Enhances Temporal Rhythmic Structure in the Envelope

Victoria Leong1, Marina Kalashnikova2, Denis Burnham2, Usha Goswami1; 1University of Cambridge, UK; 2University of Western Sydney, Australia
Thu-O-36-4

Infant-directed speech (IDS) supports language learning via mechanisms that are still not well-understood. Here, we adopt a ‘temporal sampling’ perspective to investigate whether rhythmic enhancements in the temporal structure of IDS could support...
multi-timescale neuronal oscillatory sampling of the speech signal by the infant brain. We compare natural maternal speech directed to infants at the ages of 7-, 9-, 11- and 19-months, to adult-directed speech (ADS). Speech temporal structure is analysed using a novel multi-timescale Spectral-Amplitude Modulation Phase Hierarchy (S-AMPH) model, which extracts the Stress-rate, Syllable-rate and Phoneme-rate modulations (i.e. temporal patterns). Compared to ADS, we find that IDS shows a 'stress-shifted' temporal profile. Stress-rate modulations dominate the modulation spectrum of IDS, whereas Syllable-rate modulations are dominant in ADS. Further, multi-timescale phase-synchronisation measures indicate that in IDS, Syllable-rate modulations are more synchronised to Stress-rate modulations and less synchronised to Phoneme-rate modulations. Thus, when speaking to infants, mothers pattern their syllables more regularly with prosodic stress, while allowing the phonemes within uttered syllables to vary more in timing. Accordingly, we conclude that the temporal structure of (Australian English) IDS is primarily stress-dominant, which could 'tune' the infant brain toward stress-based speech segmentation — an adaptive strategy for boot-strapping early language learning.

Influences of Tone Sandhi on Word Recognition in Preschool Children
Dilu Wewalaarachchi, Leher Singh; National University of Singapore, Singapore

Language learners have to contend with systematic variation in the input in the form of morpho-phonological change. In particular, tone Sandhi refers to a morphophonemic alternation referring to a license tone substitution under particular conditions. Two cohorts of children (3-4 years and 4-5 years) were tested on their ability to recognize words under four different conditions. The first three conditions entailed a change in the form of a word due to i) a native phonological rule (Tone 3 Sandhi), ii) a phonologically illegal tone substitution and iii) absence of a substitution when Tone 3 Sandhi was warranted. The fourth condition consisted of the correct, unaltered form of a word, which served as a control condition. Results demonstrated that 3 to 4 year old children were not able to recognize words that were subjected to Sandhi even when Sandhi was licensed by the phonological context, although they were able to correctly identify instances where the Sandhi rule was omitted, resulting in mispronunciations. By 5 years, children were able to recognize words when Sandhi was applied correctly. Results point to a stabilization of word recognition abilities during the preschool years with respect to suprasegmental morphophonemic change.

Lexical Representation of Consonant, Vowels and Tones in Early Childhood
Hwee Hwee Goh 1, Charlene Hu 1, Kheng Hui Yeo 2, Leher Singh 1, 1 National University of Singapore, Singapore; 2 A*STAR, Singapore

Several previous studies have investigated the extent to which segmental detail is represented in the developing lexicon. However, the majority of previous studies have focused on consonant and vowel representation, with little attention to the representation of lexicon tone in spite of its predominance in languages of the world. The current research provides a direct comparison of vowels, consonants and lexical tone representation using a mispronunciation paradigm in children at two ages (2.5 to 3.5 years and 4 to 5 years). Results point to asynchronous emergence of sensitivity to vowel, consonant and tone variation. Tone sensitivity appeared to be high in the younger cohort relative to vowels and consonants. By contrast, vowel and consonant sensitivity was strong in the older age group, with tone sensitivity appearing relatively weak. Findings point to a close coupling in vowel and consonant sensitivity in the preschool years and a dissociation with lexical tone.

Audiovisual Temporal Sensitivity in Typical and Dyslexic Adult Readers
Ana A. Francisco 1, Alexandra Jesse 2, Margriet A. Groen 1, James M. McQueen 1; 1 Radboud Universiteit Nijmegen, The Netherlands; 2 UMass Amherst, USA

Reading is an audiovisual process that requires the learning of systematic links between graphemes and phonemes. It is thus possible that reading impairments reflect an audiovisual processing deficit. In this study, we compared audiovisual processing in adults with developmental dyslexia and adults without reading difficulties. We focused on differences in cross-modal temporal sensitivity both for speech and for non-speech events. When compared to adults without reading difficulties, adults with developmental dyslexia presented a wider temporal window in which unsynchronized speech events were perceived as synchronized. No differences were found between groups for the non-speech events. These results suggest a deficit in dyslexia in the perception of cross-modal temporal synchrony for speech events.

Aero-Tactile Integration in Fricatives: Converting Audio to Air Flow Information for Speech Perception Enhancement
Donald Derrick, Greg A. O’Beirne, Tom De Rybel, Jennifer Hay; University of Canterbury, New Zealand

We follow up on research demonstrating that aero-tactile information can enhance or interfere with accurate auditory perception among uninformed and untrained perceivers [1, 2, 3]. We computationally extract aperiodic information from auditory recordings of speech, which represents turbulent air-flow produced from the lips [4, 5]. This extracted signal is used to drive a piezoelectric air-pump producing air-flow to the right temple simultaneous with presentation of auditory recordings. Using forced-choice experiments, we replicate previous results with stops, finding enhanced perception of /pa/ in /pa/ vs. /ba/ pairs and /ta/ in /ta/ vs. /da/ pairs [1, 6, 2, 3]. We also found enhanced perception of /la/ in /la/ vs. /ha/ pairs, and /sha/ in /da/ vs. /sha/ pairs, demonstrating that air flow during fricative production contacting the skin can also enhance speech perception. The results show that aero-tactile information can be extracted from the audio signal and used to enhance speech perception of a large class of speech sounds found in many languages of the world.

Relative Importance of AM and FM Cues for Speech Comprehension: Effects of Speaking Rate and their Implications for Neurophysiological Processing of Speech
Guangting Mai; University of Sheffield, UK

Previous studies have shown that slowly-varying amplitude modulations (AM) are crucial for speech comprehension. Moreover,
recent neurophysiological studies showed low-frequency neural oscillations (<10 Hz) are taking roles in tracking such critical AM cues which facilitate speech comprehension. However, many of such studies neglected the detailed spectral information (frequency modulations (FM)). The current paper conducted a behavioral experiment to study the relative importance of AM and FM cues for sentence intelligibility based on the hypothesis that such importance is modulated by speaking rate. By measuring the intelligibility of Mandarin sentences with selective removal of AM cues at particular AM rates or replacing FM cues with Gaussian noise, the current study found: (1) at a low speaking rate (4-Hz syllable rate), FM cues and high-rate AM cues made only marginal contributions to speech intelligibility, which is consistent with previous findings; (2) at high speaking rates (6- and 8-Hz syllable rates), however, FM cues made significant contributions even as the AM cues had a strong intelligibility advantage for the standard accent when mixed with noise, replicating previous findings; (2) at high speaking rates (6- and 8-Hz syllable rates), however, FM cues made significant contributions even as the AM cues had a strong intelligibility advantage for the standard accent when mixed with noise, replicating previous neurophysiological studies. This result thus illustrates the relative importance of AM and FM cues at different speaking rates in Mandarin and implications for the neurophysiological speech processing were further discussed.

The Effect of Regional and Non-Native Accents on Word Recognition Processes: A Comparison of EEG Responses in Quiet to Speech Recognition in Noise

Louise Stringer, Paul Iverson; University College London, UK

Thu O-37-4

Previous work has shown that both the speed and accuracy of word recognition can be reduced if a talker has a regional or non-native accent, particularly under noisy conditions. This study investigated whether the reduced intelligibility of some accents in noise are related to aspects of speech processing in quiet. Our goal was to see if difficulties processing accented speech under adverse listening conditions can be revealed in quiet using electrophysiological methods. Participants heard English sentences in a standard, regional or non-native accent. Behavioural measures found listeners had a strong intelligibility advantage for the standard accent when mixed with noise, replicating previous work. However, differences in processing the accents were less clear using the EEG measures. We found a significant Phonological Mapping Negativity (PMN) elicited by phonological anomalies, as well as an N400 effect, which is related to semantic anomalies, but greater than AM cues at particular AM rates or replacing FM cues with Gaussian noise, the current study found: (1) at a low speaking rate (4-Hz syllable rate), FM cues and high-rate AM cues made only marginal contributions to speech intelligibility, which is consistent with previous findings; (2) at high speaking rates (6- and 8-Hz syllable rates), however, FM cues made significant contributions even as the AM cues had a strong intelligibility advantage for the standard accent when mixed with noise, replicating previous neurophysiological studies. This result thus illustrates the relative importance of AM and FM cues at different speaking rates in Mandarin and implications for the neurophysiological speech processing were further discussed.

Towards a Neural Measure of Perceptual Distance — Classification of Electroencephalographic Responses to Synthetic Vowels

Manson C.-M. Fong, James W. Minett, Thierry Blu, William S.-Y. Wang; Chinese University of Hong Kong, China

Thu O-37-5

How vowels are organized cortically has previously been studied using auditory evoked potentials (AEPs), one focus of which is to determine whether perceptual distance could be inferred using AEP components. The present study extends this line of research by adopting a machine-learning framework to classify evoked responses to four synthetic mid-vowels differing only in second formant frequency (F2 = 840, 1200, 1680, and 2280 Hz). 6 subjects attended 4 EEG sessions each on separate days. Classifiers were trained using time-domain data in successive time-windows of various sizes. Results were the most accurate when a window of about 80 ms was used. By integrating the scores from individual classifiers, the maximum mean binary classification rates improved to 70% (10 trials) and 77% (20 trials). To assess how well perceptual distances among the vowels were reflected in our results, discriminability indices (d’) were computed using both the behavioral results in a screening test and the classification results. It was found that the two sets of indices were significantly correlated. The pair that was the most (least) discriminable behaviorally was also the most (least) classifiable neurally. Our results support the use of classification methodology for developing a neural measure of perceptual distance.

Collecting a Corpus of Dutch Noise-Induced ‘Slips of the Ear’

Odette Scharenborg, Eric Sanders, Bert Cranen; Radboud Universiteit Nijmegen, The Netherlands

Thu O-37-6

When trying to understand how listeners recognise words, listeners’ misperceptions, so-called ‘slips of the ear’, can reveal important aspects of the underlying mechanisms of normal word recognition. Such misperceptions shed light onto how inferences are made by listeners about acoustic details in the speech signal and how these interact with other sound sources in the background. On the other hand, if speech from a particular speaker is more prone to being misperceived than that from another speaker, these misperceptions may also shed light onto speaker characteristics. To study these phenomena, misperceptions that occur consistently are invaluable. Although such confusions are quite rare, within the Marie Curie INSPIRE project, software has been developed to efficiently collect such consistent confusions for different languages. Using this software, we have started to collect Dutch consistent confusions. Single words, embedded in five different types of noise at different SNRs, produced by four speakers were presented to Dutch listeners. In a preliminary analysis, consistent confusions were analysed in terms of phoneme substitutions, insertions, and deletions, reconstructions of words using background noise, and eccentric cases. Moreover, the number and types of consistent confusions obtained in the different noise types and from different speakers are compared.

Poster Session 25: Language and Lexical Modeling

Max Atria Gallery
10:00 – 12:00, Thursday 18 September 2014
Chair: Alexander Rudnicky

Lexical Modeling for Arabic ASR: A Systematic Approach

Tuka Al Hanai, James R. Glass; MIT, USA

Thu-P-25-1, Poster

Arabic has an ambiguous mapping between words and pronunciations, making it a deep orthographic system. This ambiguity can be resolved through diacritics, which if displayed, would compose 30% of characters in a text. We investigate the different dimensions of lexical modeling, covering diacritics, pronunciation rules, and acoustic based pronunciation modeling. We show the impact of explicitly modeling the different classes of diacritics (short vowels, geminates, nummations). We further show that a phonetic lexicon, derived by applying simple pronunciation rules to diacritized words, offers the best gains in ASR perfor-
Hybrid Language Models for Speech Transcription

Luiza Orosanu, Denis Jouvet; LORIA, France

This paper analyzes the use of hybrid language models for automatic speech transcription. The goal is to later use such an approach as a support for helping communication with deaf people, and to run it on an embedded decoder on a portable device, which introduces constraints on the model size. The main linguistic units considered for this task are the words and the syllables. Various lexicon sizes are studied by setting thresholds on the word occurrence frequencies in the training data, the less frequent words being therefore syllabified. A recognizer using this kind of language model can output between 62% and 96% of words (with respect to the thresholds on the word occurrence frequencies; the other recognized lexical units are syllables). By setting different thresholds on the confidence measures associated to the recognized words, the most reliable word hypotheses can be identified, and they have correct recognition rates between 70% and 92%.

Neural Network Language Models for Low Resource Languages

Ankur Gandhe, Florian Metze, Ian Lane; Carnegie Mellon University, USA

For resource rich languages, recent works have shown Neural Network based Language Models (NNLMs) to be an effective modeling technique for Automatic Speech Recognition, out performing standard n-gram language models (LMs). For low resource languages, however, the performance of NNLMs has not been well explored. In this paper, we evaluate the effectiveness of NNLMs for low resource languages and show that NNLMs learn better word probabilities than state-of-the-art n-gram models even when the amount of training data is severely limited. We show that interpolated NNLMs obtain a lower WER than standard n-gram models, no matter the amount of training data. Additionally, we observe that with small amounts of data (approx. 100k training tokens), feed-forward NNLMs obtain lower perplexity than recurrent NNLMs, while for the larger data condition (500k–1M training tokens), recurrent NNLMs can obtain lower perplexity than feed-forward models.

Feed Forward Pre-Training for Recurrent Neural Network Language Models

Siva Reddy Gangireddy, Ferguson McInnes, Steve Renals; University of Edinburgh, UK

The recurrent neural network language model (RNNLM) has been demonstrated to consistently reduce perplexities and automatic speech recognition (ASR) word error rates across a variety of domains. In this paper we propose a pre-training method for the RNNLM, by sharing the output weights of a feed-forward neural network language model (NNLM) with the RNNLM. This is accomplished by first fine-tuning the weights of the NNLM, which are then used to initialise the output weights of an RNNLM with the same number of hidden units. We have carried out text-based experiments on the Penn Treebank Wall Street Journal data, and ASR experiments on the TED talks data used in the International Workshop on Spoken Language Translation (IWSLT) evaluation campaigns. Across the experiments, we observe small improvements in perplexity and ASR word error rate.

NOTES
Institute Europe, Germany

features.

or hundreds of billions of words and hundreds of millions of rate reductions, even in very large scale training scenarios of tens model quality substantially, as shown in perplexity and word-error backoff cost is not required to regularize the model, we investigate models using Katz backoff [3] and similar techniques. Even though of backoff or mixture weights used in smoothed n-gram language in the likelihood objective, hence avoiding the need for the kinds models that are typically regularized via well-known L1 or L2 terms Maximum Entropy (MaxEnt) language models [1, 2] are linear improves the detection.

that the neighboring words contain information which further experiment, are suited for prominence detection and, furthermore, that sequence models such as CRF, which performs best in our comparisons.

chain Conditional Random Fields (CRF). Both were trained on

correction.

MMT model is then used in conjunction with a Pronunciation

in the speech recognition framework.

models, with the best results achieved with a recurrent neural network based language model. The baseline unpruned Kneser-Ney 5-gram model achieves perplexity 67.6. A combination of techniques leads to 35% reduction in perplexity, or 10% reduc-

tion in cross-entropy (bits), over that baseline.

The benchmark is available as a code.google.com project; besides the scripts needed to rebuild the training/held-out data, it also makes available log-probability values for each word in each of ten held-out data sets, for each of the baseline N-gram models.

Integrating Sequence Information in the Audio-Visual Detection of Word Prominence in a Human-Machine Interaction Scenario

Andrea Schnall 1, Martin Heckmann 2; 1Technische Universität Darmstadt, Germany; 2Honda Research Institute Europe, Germany

Thu-P-25-8, Poster

Modifying the articulatory parameters to raise the prominence of a segment of an utterance (hyperarticulating) is usually accompa-
nied by a reduction of these parameters (hypoarticulation) for the neighboring segments. In this paper we investigate different approaches for the automatic labeling of the prominence of words. In particular, we investigate how the information in the sequence can be used. During the recording of the underlying audio-visual database, the subjects were asked to make corrections for a misunder-

standing of a single word of the system by using prosodic cues only. We extracted an extensive range of features from the audio and visual channel. For the classification of word prominence we compare two algorithms. On the one hand SVM, a local classifier, on the other hand a classifier based on a sequential model, linear chain Conditional Random Fields (CRF). Both were trained on different context regions. For the CRF the whole sentence is used as a word sequence for training and testing. Overall we show that sequence models such as CRF, which performs best in our experiment, are suited for prominence detection and, furthermore, that the neighboring words contain information which further improves the detection.

Backoff Inspired Features for Maximum Entropy Language Models

Fadi Biadsy, Keith Hall, Pedro J. Moreno, Brian Roark; Google, USA

Thu-P-25-9, Poster

Maximum Entropy (MaxEnt) language models [1, 2] are linear models that are typically regularized via well-known L1 or L2 terms in the likelihood objective, hence avoiding the need for the kinds of backoff or mixture weights used in smoothed n-gram language models using Katz backoff [3] and similar techniques. Even though backoff cost is not required to regularize the model, we investigate the use of backoff features in MaxEnt models, as well as some backoff-inspired variants. These features are shown to improve model quality substantially, as shown in perplexity and word-error rate reductions, even in very large scale training scenarios of tens or hundreds of billions of words and hundreds of millions of features.

BioKIT — Real-Time Decoder for Biosignal Processing

Dominic Telaar, Michael Wand, Dirk Gehrig, Felix Putze, Christoph Amma, Dominic Heger, Ngoc Thang Vu, Mark Erhardt, Tim Schlippe, Matthias Janke, Christian Herff, Tanja Schultz; KIT, Germany

Thu-P-25-10, Poster

We introduce BioKIT, a new Hidden Markov Model based toolkit to preprocess, model and interpret biosignals such as speech, motion, muscle and brain activities. The focus of this toolkit is to enable researchers from various communities to pursue their experiments and integrate real-time biosignal interpretation into their applications. BioKIT boosts a flexible two-layer structure with a modular C++ core that interfaces with a Python scripting layer, to facilitate development of new applications. BioKIT employs sequence-level parallelization and memory sharing across threads. Additionally, a fully integrated error blaming component facilitates in-depth analysis. A generic terminology keeps the barrier to entry for researchers from multiple fields to a minimum. We describe our online capable dynamic decoder and report on initial experi-

ments on three different tasks. The presented speech recognition experiments employ Kaldi [1] trained deep neural networks with the results set in relation to the real time factor needed to obtain them.

Speech Recognition Without a Lexicon — Bridging the Gap Between Graphemic and Phonetic Systems

David Harwath, James R. Glass; MIT, USA

Thu-P-25-11, Poster

Modern speech recognizers rely on three core components: an acoustic model, a language model, and a pronunciation lexicon. In order to expand speech recognition capability to low-resource languages and domains, techniques to peel away the expert knowl-

dge required to craft these three components have been growing in popularity. In this paper, we present a method for automati-
cally learning a weighted pronunciation lexicon in a data-driven fashion without assuming the existence of any phonetic lexicon whatsoever. Given an initial grapheme acoustic model, our method utilizes a novel technique for semi-constrained acoustic unit decoding, which is used to help train a letter to sound (L2S) model. The L2S model is then used in conjunction with a Pronunciation Mixture Model (PMM) to infer a pronunciation lexicon. We evaluate our method on English as well as Lao and Haitian, two low-resource languages featured in the IARPA Babel program.

Poster Session 26: Speech Enhancement (Single- and Multi-Channel)

Max Atria Gallery
10:00 - 12:00, Thursday 18 September 2014
Chair: Elmar Nöth

A New Auxiliary-Vector Algorithm with Conjugate Orthogonality for Speech Enhancement

Shengkui Zhao 1, Douglas L. Jones 2; 1ADSC, Singapore; 2University of Illinois at Urbana-Champaign, USA

Thu-P-26-1, Poster

In this paper, we propose a new auxiliary-vector (AV) algorithm using the conjugate orthogonality for speech enhancement. When only a limited data record is available, the AV algorithm is the state-of-the-art for obtaining the minimum-variance-distortionless (MVDR) filter. However, the current AV algorithms suffer from convergence problems when applied to the speech enhancement.

NOTES
Based on the conjugate Gram-Schmidt process, we develop new auxiliary vectors that are conjugate orthogonal and apply them to the AV algorithm. The proposed conjugate AV algorithm converges to the optimal MVDR solution within finite steps no greater than the filter dimension. Theoretical analysis establishes formal convergence of the proposed conjugate AV algorithm. Our experiments using the synthetic and real speech data show favorites of the new proposal over the state-of-the-art approaches.

Acoustic Characteristics of Critical Message Utterances in Noise Applied to Speech Intelligibility Enhancement

Neehar Jathar, Preeti Rao; IIT Bombay, India

Motivated by the potential of speech modification for the enhancement of intelligibility in noisy environments, we study the acoustic characteristics of speech produced in the context of critical announcements made in noisy listening situations. A corpus of 3 speakers producing 20 Marathi train station announcements is analysed for articulatory-acoustic and prosodic differences between speech in noise and in quiet. It is observed that apart from the global changes that are characteristic of increased vocal effort, the intonation associated with phrase and word accents is consistently modified. Listening tests with modified speech suggest that spectral shaping and F0 modifications that are linguistically valid contribute constructively to increased intelligibility in noise, as measured in an information extraction task.

Dynamic Noise Aware Training for Speech Enhancement Based on Deep Neural Networks

Yong Xu¹, Jun Du¹, Li-Rong Dai², Chin-Hui Lee²;¹USTC, China; ²Georgia Institute of Technology, USA

We propose three algorithms to address the mismatch problem in deep neural network (DNN) based speech enhancement. First, we investigate noise aware training by incorporating noise information in the test utterance with an ideal binary mask based dynamic noise estimation approach to improve DNN’s speech separation ability from the noisy signal. Next, a set of more than 100 noise types is adopted to enrich the generalization capabilities of the DNN to unseen and non-stationary noise conditions. Finally, the quality of the enhanced speech can further be improved by global variance equalization. Empirical results show that each of the three proposed techniques contributes to the performance improvement. Compared to the conventional logarithmic minimum mean squared error speech enhancement method, our DNN system achieves 0.32 PESQ (perceptual evaluation of speech quality) improvement across six signal-to-noise ratio levels ranging from -5dB to 20dB on a test set with unknown noise types. We also observe that the combined strategies can well suppress highly non-stationary noise better than all the competing state-of-the-art techniques we have evaluated.

Microphone Array Post-Filtering Using Supervised Machine Learning for Speech Enhancement

Pasi Pirttila, Joonas Nikunen; Tampere University of Technology, Finland

High level of noise reduces the perceptual quality and intelligibility of speech. Therefore, enhancing the captured speech signal is important in everyday applications such as telephony and teleconferencing. Microphone arrays are typically placed at a distance from a speaker and require processing to enhance the captured signal. Beamforming provides directional gain towards the source of interest and attenuation of interference. It is often followed by a single channel post-filter to further enhance the signal. Non-linear spatial post-filters are capable of providing high noise suppression but can produce unwanted musical noise that lowers the perceptual quality of the output. This work proposes an artificial neural network (ANN) to learn the structure of naturally occurring post-filters to enhance speech from interfering noise. The ANN uses phase-based features obtained from a multichannel array as an input. Simulations are used to train the ANN in a supervised manner. The performance is measured with objective scores from speech recorded in an office environment. The post-filters predicted by the ANN are found to improve the perceptual quality over delay-and-sum beamforming while maintaining high suppression of noise characteristic to spatial post-filters.

Novel Speech Duration Modifier for Packet Based Communication System

Senthil Kumar Mani, Jitenendra Kumar Dhiman, K. Sri Ramu Maruty; IIT Hyderabad, India

In this paper, we propose a real-time method for duration modification of speech for packet based communication system. While there is rich literature available on duration modification, it fails to clearly address the issues in real-time implementation of the same. Most of the duration modification methods rely on accurate estimation of pitch marks, which is not feasible in a real-time scenario. The proposed method modifies the duration of Linear Prediction residual of individual frames without using any look-ahead delay and knowledge of pitch marks. In this method, multiples of pitch period is repeated or removed from a frame depending on a scheduling algorithm. The subjective quality of the proposed method was found to be better than waveform similarity overlap and add (WSOLA) technique as well as Linear Prediction Synchronous Overlap and Add (LP-PSOLA) technique.

Experiments on Deep Learning for Speech Denoising

Ding Liu, Paris Smaragdis, Minje Kim; University of Illinois at Urbana-Champaign, USA

In this paper we present some experiments using a deep learning model for speech denoising. We propose a very lightweight procedure that can predict clean speech spectra when presented with noisy speech inputs, and we show how various parameter choices impact the quality of the denoised signal. Through our experiments we conclude that such a structure can perform better than some comparable single-channel approaches and that it is able to generalize well across various speakers, noise types and signal-to-noise ratios.

Single-Channel Dynamic Exemplar-Based Speech Enhancement

Nasser Mohammadiha, Simon Doclo; Carl von Ossietzky Universität Oldenburg, Germany

This paper proposes an exemplar-based speech enhancement method based on high-resolution STFT magnitude spectrograms, where a selection of the nonnegative training data is used as the dictionary to provide a holistic nonnegative representation of the test data. We discuss how this exemplar-based model ensures that the enhanced speech signal falls on the speech manifold, which improves the quality of the enhanced speech signal. To
explore the temporal continuity, a vector autoregressive model is used to model the activations where the model parameters are learned using a new NMF-based approach. Results from several supervised and semi-supervised speech enhancement experiments indicate that the proposed exemplar-based method outperforms the considered supervised and unsupervised denoising algorithms in terms of both segmental SNR and PESQ at different input SNRs.

Using Hidden Markov Models for Speech Enhancement

Akhihiro Kato, Ben Milner; University of East Anglia, UK

This work presents an approach to speech enhancement that operates using a speech production model to reconstruct a clean speech signal from a set of speech parameters that are estimated from the noisy speech. The motivation is to remove the distortion and residual and musical noises that are associated with conventional filtering-based methods of speech enhancement. The STRAIGHT vocoder forms the model for speech reconstruction and requires a time-frequency surface and fundamental frequency information. Hidden Markov model synthesis is used to estimate an equation of the time-frequency surface and this is combined with the noisy surface using a perceptually motivated signal-to-noise ratio weighting. Experimental results compare the proposed reconstruction-based method to conventional filtering-based approaches of speech enhancement.

Blind Source Extraction Based on a Direction-Dependent a-priori SNR

Lukas Pfeifenberger, Franz Pernkopf; Technische Universität Graz, Austria

In many hands-free applications, we encounter a speaker located in the near-field embedded in diffuse far-field noise. In this paper, we contribute an algorithm to estimate the speech and noise power spectral density (PSD) based on a direction-dependent SNR (DD-SNR). The only prior knowledge needed is a model of the diffuse noise sound field. The enhanced speech signal is obtained by a parametric multi-channel Wiener filter (PMWF), which is constructed without any speech presence or absence probabilities, or smoothing in frequency. We achieve high speech quality and sufficient noise reduction by iteratively improving the speech PSD estimate using the output of the PMWF. The performance of our algorithm is demonstrated by using the PESQ and PEASS measures.

Least Squares Phase Estimation of Mixed Signals

Carlos Eduardo Cancino Chacón, Pejman Mowlaee; Technische Universität Graz, Austria

Estimating the phase of sinusoids in noise in contrast to amplitude and frequency components has been less addressed in previous studies. In this paper, we derive the least squares phase estimator (LSPE) solution to recover the phase of an underlying signal observed in noise. Through Monte-Carlo simulations, we demonstrate the robustness of the proposed phase estimator against the modeling error. The proposed phase estimator is further evaluated in the speech enhancement setup to assess how much improvement is obtained by replacing the noisy phase with the LSPE when reconstructing the enhanced speech signal. Significant improvement in speech quality and speech intelligibility is obtained by replacing the noisy phase with the estimated phase provided by the proposed LSPE once the ambiguity in the phase candidates is removed.

Speech Enhancement from Additive Noise and Channel Distortion — A Corpus-Based Approach

Ji Ming, Danny Crookes; Queen’s University Belfast, UK

This paper presents a new approach to single-channel speech enhancement involving both noise and channel distortion (i.e., convolutional noise). The approach is based on finding longest matching segments (LMS) from the corpus of clean, wideband speech. The approach adds three novel developments to our previous LMS research. First, we address the problem of channel distortion as well as additive noise. Second, we present an improved method for modeling noise. Third, we present an iterative algorithm for improved speech estimates. In experiments using speech recognition as a test with the Aurora 4 database, the use of our enhancement approach as a preprocessor for feature extraction significantly improved the performance of a baseline recognition system. In another comparison against conventional enhancement algorithms, both the PESQ and the segmental SNR ratings of the LMS algorithm were superior to the other methods for noisy speech enhancement.

Poster Session 27: Robust ASR

Max Atria Gallery
10:30 - 12:00, Thursday 18 September 2014
Chair: Alexandros Potamianos

Robust Speech Recognition in Reverberant Environments Using Subband-Based Steady-State Monaural and Binaural Suppression

Hyung-Min Park 1, Matthew Maciejewski 2, Chanwoo Kim 3, Richard M. Stern 2; 1Sogang University, Korea; 2Carnegie Mellon University, USA; 3Google, USA

The precedence effect describes the ability of the auditory system to suppress the later-arriving components of sound in a reverberant environment, maintaining the perceived arrival azimuth of a sound in the direction of the actual source, even though later reverberant components may arrive from other directions. It is also widely believed that precedence-like processing can also improve speech intelligibility, as well as the accuracy of speech recognition systems, in reverberant environments. While the mechanisms underlying the precedence effect have traditionally been assumed to be binaural in nature, it is also possible that the suppression of later components may take place monaurally, and that the suppression of the later-arriving components of the spatial image may be a consequence of this more peripheral processing. This paper compares the potential contributions of onset enhancement (and consequent steady-state suppression) of the envelopes of subband components of speech at both the monaural and binaural levels. Experimental results indicate that substantial improvement in recognition accuracy can be obtained in reverberant environments if the feature extraction includes both onset enhancement and binaural interaction. Recognition accuracy appears to be relatively unaffected by the stage in the suppression processing at which the binaural interaction takes place.

Variable-Component Deep Neural Network for Robust Speech Recognition

Rui Zhao 1, Jinyu Li 2, Yifan Gong 2; 1Microsoft, China; 2Microsoft, USA

In this paper, we propose variable-component DNN (VCDNN) to...
improve the robustness of context-dependent deep neural network hidden Markov model (CD-DNN-HMM). This method is inspired by the idea from variable-parameter HMM (VPHMM) in which the variation of model parameters are modeled as a set of polynomial functions of environmental signal-to-noise ratio (SNR), and during the testing, the model parameters are recomputed according to the estimated testing SNR. In VCDNN, we refine two types of DNN components: (1) weighting matrix and bias (2) the output of each layer. Experimental results on Aurora4 task show VCDNN achieved 6.5% and 5.92% relative word error rate reduction (WERR) over the standard DNN for the two methods, respectively. Under unseen SNR conditions, VCDNN gave even better result (8.46% relative WERR for the DNN varying matrix and bias, 7.08% relative WERR for the DNN varying layer output). Moreover, VCDNN with 1024 units per hidden layer beats the standard DNN with 2048 units per hidden layer with 3.22% WERR and a half computational/memory cost reduction, showing superior ability to produce sharper and more compact models.

Effective Modulation Spectrum Factorization for Robust Speech Recognition
Yu-Chen Kao, Yi-Ting Wang, Berlin Chen; National Taiwan Normal University, Taiwan
Thu-P-27-3, Poster

Modulation spectrum processing of acoustic features has received considerable attention in the area of robust speech recognition because of its relative simplicity and good empirical performance. An emerging school of thought is to conduct nonnegative matrix factorization (NMF) on the modulation spectrum domain so as to distill intrinsic and noise-invariant temporal structure characteristics of acoustic features for better robustness. This paper presents a continuation of this general line of research and its main contribution is two-fold. One is to explore the notion of sparsity for NMF so as to ensure the derived basis vectors have sparser and more localized representations of the modulation spectra. The other is to investigate a novel cluster-based NMF processing, in which speech utterances belonging to different clusters will have their own set of cluster-specific basis vectors. As such, the speech utterances can retain more discriminative information in the NMF processed modulation spectra. All experiments were conducted on the Aurora-2 corpus and task. Empirical evidence reveals that our methods can offer substantial improvements over the baseline NMF method and achieve performance competitive to or better than several widely-used robustness methods.

Hybrid MLP/Structured-SVM Tandem Systems for Large Vocabulary and Robust ASR
Suman V. Ravuri; ICISI, USA
Thu-P-27-4, Poster

Tandem systems based on multi-layer perceptrons (MLPs) have improved the performance of automatic speech recognition systems on both large vocabulary and noisy tasks. One potential problem of the standard Tandem approach, however, is that the MLPs generally used do not model temporal dynamics inherent in speech. In this work, we propose a hybrid MLP/Structured-SVM model, in which the parameters between the hidden layer and output layer and temporal transitions between output layers are modeled by a Structured-SVM. A Structured-SVM can be thought of as an extension to the classical binary support vector machine which can naturally classify “structures” such as sequences. Using this approach, we can identify sequences of phones in an utterance. We try this model on two different corpora — Aurora2 and the large-vocabulary section of the ICSI meeting corpus — to investigate the model’s performance in noisy conditions and on a large-vocabulary task. Compared to a difficult Tandem baseline in which the MLP is trained using 2nd-order optimization methods, the MLP/Structured-SVM system decreases WER in noisy conditions by 7.9% relative. On the large vocabulary corpus, the proposed system decreases WER by 1.1% absolute compared to the 2nd-order Tandem system.

Robust Speech Recognition Using Temporal Masking and Thresholding Algorithm
Chanwoo Kim 1, Keon K. Chin 1, Michiel Bacchiani 1, Richard M. Stern 2; 1 Google, USA; 2 Carnegie Mellon University, USA
Thu-P-27-5, Poster

In this paper, we present a new dereverberation algorithm called Temporal Masking and Thresholding (TMT) to enhance the temporal spectra of spectral features for robust speech recognition in reverberant environments. This algorithm is motivated by the precedence effect and temporal masking of human auditory perception. This work is an improvement of our previous dereverberation work called Suppression of Slowly-varying components and the falling edge of the power envelope (SSF). The TMT algorithm uses a different mathematical model to characterize temporal masking and thresholding compared to the model that had been used to characterize the SSF algorithm. Specifically, the nonlinear highpass filtering used in the SSF algorithm has been replaced by a masking mechanism based on a combination of peak detection and dynamic thresholding. Speech recognition results show that the TMT algorithm provides superior recognition accuracy compared to other algorithms such as LTLS, VTS, or SSF in reverberant environments.

Deep Neural Network Bottleneck Features for Generalized Variable Parameter HMMs
Xurong Xie, Rongfeng Su, Xuning Liu, Lan Wang; Chinese Academy of Sciences, China
Thu-P-27-6, Poster

Recently deep neural networks (DNNs) have become increasingly popular for acoustic modelling in automatic speech recognition (ASR) systems. As the bottleneck features they produce are inherently discriminative and contain rich hidden factors that influence the surface acoustic realization, the standard approach is to augment the conventional acoustic features with the bottleneck features in a tandem framework. In this paper, an alternative approach to incorporate bottleneck features is investigated. The complex relationship between acoustic features and DNN bottleneck features is modelled using generalized variable parameter HMMs (GVP-HMMs). The optimal GVP-HMM structural configuration and model parameters are automatically learnt. Significant error rate reductions of 48% and 8% relative were obtained over the baseline multi-style HMM and tandem HMM systems respectively on Aurora 2.

A Novel Dynamic Parameters Calculation Approach for Model Compensation
Suliang Bu, Yanmin Qian, Kai Yu; Shanghai Jiao Tong University, China
Thu-P-27-7, Poster

Model compensation approach has been successfully applied to various noise robust speech recognition tasks. In this paper, based on Continuous Time (CT) approximation, the dynamic mismatch function is derived without further approximation. With such mismatch function, a novel approach to deriving the formula for
calculating the dynamic statistics is presented. Besides, we also provide an insight on the processing of the pseudo inverse of non-square discrete cosine transform (DCT) matrix during model compensation. Experiments on Aurora 4 showed that the proposed approach obtained 23.2% relative WER reduction over traditional first-order Vector Taylor Series (VTS) approach.

Speech Recognition Based on Itakura-Saito Divergence and Dynamics/Sparseness Constraints from Mixed Sound of Speech and Music by Non-Negative Matrix Factorization

Naoaki Hashimoto, Shoichi Nakano, Kazumasa Yamamoto, Seiichi Nakagawa; Toyohashi University of Technology, Japan

We considered a speech recognition method for mixed sound, which is composed of both speech and music, that only removes music based on non-negative matrix factorization (NMF). We used Itakura-Saito divergence instead of Kullback-Leibler divergence to compare the cost function, and the dynamics and sparseness constraints of a weight matrix to improve speech recognition. For isolated word recognition using the matched condition model, we reduced the word error rate by 52.1% relative from the case that didn’t remove music (on average, from 69.3% to 85.3%).

Noise Robust Speech Recognition Based on Noise-Adapted HMMs Using Speech Feature Compensation

Yong-Joo Chung; Keimyung University, Korea

In conventional VTS-based noisy speech recognition methods, the parameters of the clean speech HMM are adapted to test noisy speech, or the original clean speech is estimated from the test noisy speech. However, in noisy speech recognition, improved performance is generally expected by employing noisy acoustic models produced by methods such as Multi-condition Training (MTR) and Multi-Model based Speech Recognition (MMSR) framework compared with using clean HMMs. Motivated by this idea, a method has been developed that can make use of the noisy acoustic models in the VTS algorithm where additive noise was adapted for the speech feature compensation. In this paper, we modified the previous method to adapt channel noise as well as additive noise. The proposed method was applied to noise-adapted HMMs trained by the MTR and MMSR and could reduce the relative word error rate by 6.5% and 7.2%, respectively, in the noisy speech recognition experiments on the Aurora 2 database.

Noise Spectrum Estimation Using Gaussian Mixture Model-Based Speech Presence Probability for Robust Speech Recognition

M.J. Alam¹, Patrick Kenny¹, Pierre Dumouchel², Douglas O'Shaughnessy³, CRIM, Canada; ²École de Technologie Supérieure, Canada; ³INRS-EMT, Canada

This work presents a noise spectrum estimator based on the Gaussian mixture model (GMM)-based speech presence probability (SPP) for robust speech recognition. Estimated noise spectrum is then used to compute a subband a posteriori signal-to-noise ratio (SNR). A sigmoid shape weighting rule is formed based on this subband a posteriori SNR to enhance the speech spectrum in the auditory domain, which is used in the Mel-frequency cepstral coefficient (MFCC) framework for robust feature, denoted here as Robust MFCC (RMFCC) extraction. The performance of the GMM-SPP noise spectrum estimator-based RMFCC feature extractor is evaluated in the context of speech recognition on the AURORA-4 continuous speech recognition task. For comparison we incorporate six existing noise estimation methods into this auditory domain spectrum enhancement framework. The ETSI advanced front-end (ETSI-AFE), power normalized cepstral coefficients (PNCC), and robust compressive gammachirp cepstral coefficients (RCGCC) are also considered for comparison purposes. Experimental speech recognition results show that, in terms of word accuracy, RMFCC provides an average relative improvements of 8.1%, 6.9% and 6.6% over RCGCC, ETSI-AFE, and PNCC, respectively. With GMM-SPP-based noise estimation method an average relative improvement of 3.6% is obtained over other six noise estimation methods in terms of word recognition accuracy.

Oral Session 38: Spoken Term Detection for Low-Resource Languages

William Hartmann¹, Viet-Bac Le², Abdel Messaoudi², Lori Lamel¹, Jean-Luc Gauvain¹; ¹LIMSI, France; ²Vocapia Research, France

Comparing Decoding Strategies for Subword-Based Keyword Spotting in Low-Resourced Languages

Min Ma¹, Justin Richards¹, Victor Soto², Julia Hirschberg², Andrew Rosenberg¹; ¹CUNY Graduate Center, USA; ²Columbia University, USA

For languages with limited training resources, out-of-vocabulary (OOV) words are a significant problem, both for transcription and keyword spotting. This paper investigates the use of subword lexical units for keyword spotting. Three strategies for using the sub-word units are explored: 1) converting word-based lattices to subword lattices after decoding, 2) performing a separate decoding for each subword type, and 3) a single decoding using all possible subword units. In these experiments, the best performance is achieved by carrying out a separate decoding for each subword type. Further gains are attained through system combination. We also find that ignoring word boundaries improves the detection of OOV keywords without significantly impacting in-vocabulary keyword detection. Results are presented on four languages from the IARPA Babel Program (Haitian Creole, Assamese, Bengali, and Zulu).

Strategies for Rescoring Keyword Search Results Using Word-Burst and Acoustic Features

Min Ma¹, Justin Richards¹, Victor Soto², Julia Hirschberg², Andrew Rosenberg¹; ¹CUNY Graduate Center, USA; ²Columbia University, USA

The identification of keyword queries in speech data from low-resource languages poses a challenge for current methods as speech recognition algorithms lack sufficient training data to produce high accuracy transcript. To compensate for these shortcomings, we extract signals from the data that are useful in keyword identification but are not being used by the speech recognizer. These signals take multiple forms — word burstiness, rescored confusion network posteriori and acoustic/prosodic qualities. The former denotes the tendency for keywords to occur in bursts within a conversational topic. We employ three different strategies to exploit this information: 1) a four-way classification of keyword hypotheses that targets low-scoring correct hits and
high-scoring false alarms, 2) ranking algorithms, and 3) a direct adjustment of keyword hit scores based on hypothesized repetition. We find that interpolating the results of these three strategies in an ensemble provides a reliable way to improve the results of keyword search.

**Word-Based Probabilistic Phonetic Retrieval for Low-Resource Spoken Term Detection**

Di Xu, Florian Metze; Carnegie Mellon University, USA

Two problems make Spoken Term Detection (STD) particularly challenging under low-resource conditions: the low quality of speech recognition hypotheses, and a high number of out-of-vocabulary (OOV) words. In this paper, we propose an intuitive way to handle OOV terms for STD on word-based Confusion Networks using phonetic similarities, and generalize it into a probabilistic and vocabulary-independent retrieval framework. We then reflect on how several heuristics and Machine Learning based methods can be incorporated into this framework to improve retrieval performance. We present experimental results on several low-resource languages from IARPA’s Babel program, such as Assamese, Bengali, Haitian, and Lao.

**A Keyword-Boosted sMBR Criterion to Enhance Keyword Search Performance in Deep Neural Network Based Acoustic Modeling**

I-Fan Chen¹, Nancy F. Chen², Chin-Hui Lee¹; ¹Georgia Institute of Technology, USA; ²A*STAR, Singapore

We propose a keyword-boosted state-level minimum Bayes risk (sMBR) criterion for training DNN-HMM hybrid keyword search systems by enhancing acoustic detail of a given list of target keyword terms. The rationale behind the proposed discriminative training strategy is to place more acoustic modeling emphasis on states appearing in the given keywords. We observed a relative gain of 1.7–6.1% in actual term weighted value (ATWV) performance with the proposed keyword-boosted sMBR training over the conventional sMBR systems when tested on the IARPA Babel program’s Vietnamese limited-language-pack task. A detailed result analysis suggests that the proposed sMBR objective function effectively improves the ATWV scores by boosting the probability of detecting keywords appearing in the system output with an increased correct and insertion rates in the decoded lattices.

**Combination of FST and CN Search in Spoken Term Detection**

Justin Chiu¹, Yun Wang¹, Jan Trmal², Daniel Povey², Guoguo Chen², Alexander I. Rudnicky¹; ¹Carnegie Mellon University, USA; ²Johns Hopkins University, USA

Spoken Term Detection (STD) focuses on finding instances of a particular spoken word or phrase in an audio corpus. Most STD systems have a two-step pipeline, ASR followed by search. Two approaches to search are common, Confusion Network (CN) based search and Finite State Transducer (FST) based search. In this paper, we examine combination of these two different search approaches, using the same ASR output. We find that the CN search performs better on shorter queries, and FST search performs better on longer queries. By combining the different search results from the same ASR decoding, we achieve better performance compared to either search approach on its own. We also find that this improvement is additive to the usual combination of decoder results using different modeling techniques.

**Low-Resource Open Vocabulary Keyword Search Using Point Process Models**

Chunxu Liu, Aren Jansen, Guoguo Chen, Keith Kintzley, Jan Trmal, Sanjeev Khudanpur; Johns Hopkins University, USA

The point process model (PPM) for keyword search is a whole-word parametric modeling framework based on the timing of phonetic events rather than the evolution of frame-level phonetic likelihoods. Recent progress in PPM training and decoding algorithms has yielded state-of-the-art phonetic search performance in high-resource settings, both in terms of accuracy and computational efficiency. In this paper, we consider PPM application to low-resource settings where the amount of transcribed speech is severely limited and the pronunciation dictionary is incomplete. By using (i) state-of-the-art deep neural network acoustic models to generate phonetic events and (ii) grapheme-to-phoneme conversion to generate pronunciations for out-of-vocabulary (OOV) keywords, we find the PPM system reaches state-of-the-art OOV search performance at a small computational cost. Moreover, due to their complementary methodologies, combining PPM outputs with the LVCSR baseline produces average relative ATWV improvements of 7% and 50% for in-vocabulary and OOV keywords, respectively (16% overall).

**Oral Session 39: Speech Coding and Transmission**

**Decorrelated Innovative Codebooks for ACELP Using Factorization of Autocorrelation Matrix**

Tom Bäckström, Christian R. Helmrich; FAU Erlangen-Nürnberg, Germany

Modern speech codecs based on Code Excited Linear Prediction (CELP) employ an analysis-by-synthesis optimization loop to find the best quantization of the source model parameters. With this approach, optimal quantization can be achieved only with an exhaustive search. Instead, we propose to use matrix factorization to decorrelate the objective function of the optimization problem, whereby the computationally expensive iteration can be avoided and optimal performance is guaranteed. We compare two factorizations of the autocorrelation matrix, the eigenvalue decomposition and Vandermonde factorization. Our experiments show that decorrelation improves perceptual SNR and gives a large reduction in computational complexity, mostly without significant impact on subjective quality.

**Stress and Accent Transmission In HMM-Based Syllable-Context Very Low Bit Rate Speech Coding**

Milos Cernak, Alexandros Lazaridis, Philip N. Garner, Petr Motlicek; Idiap Research Institute, Switzerland

In this paper, we propose a solution to reconstruct stress and accent contextual factors at the receiver of a very low bitrate speech codec built on recognition/synthesis architecture. In speech synthesis, accent and stress symbols are predicted from the text, which is not available at the receiver side of the speech codec. Therefore, speech signal-based symbols, generated as

---

**NOTES**
syllable-level log average F0 and energy acoustic measures, quantized using a scalar quantization, are used instead of accentual and stress symbols for HMM-based speech synthesis. Results from incremental real-time speech synthesis confirmed, that a combination of F0 and energy signal-based symbols can replace their counterparts of text-based binary accent and stress symbols developed for text-to-speech systems. The estimated transmission bit-rate overhead is about 14 bits/second per acoustic measure.

Subjective Voice Quality Evaluation of Artificial Bandwidth Extension: Comparing Different Audio Bandwidths and Speech Codecs

Hannu Pulakka1, Anssi Rämö2, Ville Myllylä3, Henri Toukoma4, Paaavo Alku4, 1Microsoft, Finland; 2Nokia Research Center, Finland; 3Aalto University, Finland
Thru-O-39.3

Artificial bandwidth extension (ABE) methods have been developed to improve the quality and intelligibility of telephone speech. In many previous studies, however, the evaluation of ABE has not fully reflected the use of ABE in mobile communication (e.g., evaluation with clean speech without coding). In this study, the subjective quality of ABE was evaluated with absolute category rating (ACR) tests involving both clean and noisy speech, two cutoff frequencies of highpass filtering, and input encoded at different bit rates. Three ABE methods were evaluated, two for narrowband-to-wideband extension and one for wideband-to-superwideband extension. Several speech codecs with different audio bandwidths were included in the tests. Narrowband-to-wideband ABE methods were found to significantly improve the speech quality when no background noise was present, and the mean quality scores were slightly but not significantly increased for noisy speech. Wideband-to-superwideband ABE also showed significant improvement in certain conditions with no background noise. ABE did not cause significant decrease of the mean scores in any of the tests.

Subjective Voice Quality Evaluation of Artificial Bandwidth Extension: Comparing Different Audio Bandwidths and Speech Codecs

Hannu Pulakka1, Anssi Rämö2, Ville Myllylä3, Henri Toukoma4, Paaavo Alku4, 1Microsoft, Finland; 2Nokia Research Center, Finland; 3Aalto University, Finland
Thru-O-39.3

Artificial bandwidth extension (ABE) methods have been developed to improve the quality and intelligibility of telephone speech. In many previous studies, however, the evaluation of ABE has not fully reflected the use of ABE in mobile communication (e.g., evaluation with clean speech without coding). In this study, the subjective quality of ABE was evaluated with absolute category rating (ACR) tests involving both clean and noisy speech, two cutoff frequencies of highpass filtering, and input encoded at different bit rates. Three ABE methods were evaluated, two for narrowband-to-wideband extension and one for wideband-to-superwideband extension. Several speech codecs with different audio bandwidths were included in the tests. Narrowband-to-wideband ABE methods were found to significantly improve the speech quality when no background noise was present, and the mean quality scores were slightly but not significantly increased for noisy speech. Wideband-to-superwideband ABE also showed significant improvement in certain conditions with no background noise. ABE did not cause significant decrease of the mean scores in any of the tests.

Stereo Acoustic Echo Suppression Using Widely Linear Filtering in the Frequency Domain

Zhong-Hua Fu, Lei Xie; Northwestern Polytechnical University, China
Thu-O-39.4

This paper proposes a stereo acoustic echo suppression method for duplex stereo teleconferencing. Firstly the stereo signals are combined in the frequency domain to build a single-channel signal. Secondly, with the single-channel signal model, an optimal Linearly Constrained Minimum Variance (LCMV) filter based on the Widely Linear theory is derived, which is applied on microphone signal directly to maximumly suppress the acoustic echo and preserve the quality of near-end signal. To estimate the normalized correlation vectors required in the LCMV filter, we employ an initial guess method, which doesn’t require double-talk detector and inter-channel pre-decorrelation. The experimental results verify that this method can preserve the near-end speech quality as well as the spatial information, and suppresses echo greatly.


Bong-Ki Lee, Inyoung Hwang, Jihwan Park, Joon-Hyuk Chang; Hanyang University, Korea
Thu-O-39.5

In this paper, we propose an enhanced adaptive muting method using a sigmoid function, which is based on a parameter tracking technique for the packet loss concealment algorithm of ITU-T G.722 speech codec. The packet loss concealment algorithm performs an adaptive muting to prevent the generation of unnecessary noises or clicks during packet loss recovery. While a conventional muting method applies the sigmoid function to the muting curve and the principal parameters of the sigmoid function are obtained by using a grid search-based training method, in the proposed muting algorithm, the parameters are substantially obtained from the previous good frames using the steepest descent algorithm, which minimizes the error between the desired signal and the reconstructed signal. From experimental results, the proposed adaptive muting method turns out to improve the performance of the conventional muting method under various experimental conditions.

A Robust Step-Size Control Algorithm for Frequency Domain Acoustic Echo Cancellation

Chao Wu, Kaiyu Jiang, Yanneng Guo, Qiang Fu, Yonghong Yan; Chinese Academy of Sciences, China
Thru-O-39.6

The presence of near-end interferers and echo path changes make it essential for an adaptive filter to vary its learning rate according to corresponding conditions. In this paper, a robust step-size control algorithm which is based on the optimization of the square of the bin-wise a posteriori error is proposed. To prevent the adaptive filter from diverging in the presence of interferences, constraints on the filter update are applied. The learning rate expression is derived and then we extend the method to multidelay block frequency domain adaptive filter (MDF) so as to meet the demand of low delay in practical application. An updating strategy for the constraints is proposed as well. Experiments are carried out to demonstrate the superiority of the proposed approach, especially in double-talk and echo path change situations.

Oral Session 40: Speech Enhancement

Peridot 204-205
13:30 – 15:30, Thursday 18 September 2014
Chairs: Dorothea Kolossa and Yifan Gong

Multi-Channel Speech Enhancement Using Sparse Coding on Local Time-Frequency Structures

Zhiyuan Zhou1, Zhaogui Ding1, Weifeng Li1, Zhiyong Wu1, Longbiao Wang2, Qingmin Liao1, 1Tsinghua University, China; 2Nagaoka University of Technology, Japan
Thu-O-40.1

A novel multi-channel speech enhancement technique is proposed in the present paper. We focus on the local sparsities of speech signals in contrast to the conventional beamforming and blind source separation methods. The technique utilizes the difference of local structures in temporary-frequency domain between the target speech and interfering signals for enhancing the target speech. We first estimate the local structures of the speech and noise signals at each time-frequency bin to form a local dictionary, and then recover the clean speech via sparse coding. The proposed algorithm is simple to implement and requires no prior knowledge of speech and noise. Our experimental evaluations demonstrate that the proposed method can suppress interferer and meantime preserve target speech more than some conventional methods.

Notes
Room reverberation is a primary cause of failure in distant speech recognition (DSR) systems. In this study, we present a multichannel spectrum enhancement method for reverberant speech recognition, which is an extension of a single-channel dereverberation algorithm based on convolutive nonnegative matrix factorization (NMF). The generalization to a multichannel scenario is shown to be a special case of convolutive nonnegative tensor factorization (NTF). The presented algorithm integrates information from across different channels in the magnitude short time Fourier transform (STFT) domain. By doing so, it eliminates any limitations on the array geometry or a need for information concerning the source location, making the algorithm particularly suitable for distributed microphone arrays. Experiments are performed on speech data using actual room impulse responses from AIR database. Relative WER improvements using a clean-trained ASR system vary from +7.1\% to +30.1\% based on the number of channels and the source to microphone distances (1 to 3 meters).

Speech Enhancement by Low-Rank and Convolutive Dictionary Spectrogram Decomposition
Zhao Chen, Brian McFee, Daniel P.W. Ellis; Columbia University, USA
Thu-O-40-3

A successful speech enhancement system requires strong models for both speech and noise to decompose a mixture into the most likely combination. However, if the noise encountered differs significantly from the system’s assumptions, performance will suffer. In previous work, we proposed a speech enhancement framework based on decomposing the noisy spectrogram into low rank background noise and a sparse activation of pre-learned templates, which requires few assumptions about the noise and showed promising results. However, when the noise is highly non-stationary or has large amplitude, the local SNR of the noisy speech can change drastically, resulting in less accurate decompositions between foreground speech and background noise. In this work, we extend the previous model by changing the modeling of the speech part to be the convolution of a sparse activation and pre-learned template patches, which enforces continuous structure within the speech and leads to better results in highly corrupted noisy mixtures.

Multiple-Order Non-Negative Matrix Factorization for Speech Enhancement
Xabier Jaureguiberry\(^1\), Emmanuel Vincent\(^2\), Gaël Richard\(^1\); \(^1\)LTCI, France; \(^2\)INRIA, France
Thu-O-40-4

Amongst the speech enhancement techniques, statistical models based on Non-negative Matrix Factorization (NMF) have received great attention. In a single channel configuration, NMF is used to describe the spectral content of both the speech and noise sources. As the number of components can have a crucial influence on separation quality, we here propose to investigate model order selection based on the variational Bayesian approximation to the marginal likelihood of models of different orders. To go further, we propose to use model averaging to combine several single-order NMFs and we show that a straightforward application of model averaging principles is inefficient as it turned out to be equivalent to model selection. We thus introduce a parameter to control the entropy of the model order distribution which makes the averaging effective. We also show that our probabilistic model nicely extends to a multiple-order NMF model where several NMFs are jointly estimated and averaged. Experiments are conducted on real data from the CHiME challenge and give an interesting insight on the entropic parameter and model order priors. Separation results are also promising as model averaging outperforms single-order model selection. Finally, our multiple-order NMF shows an interesting gain in computation time.

NMF-Based Speech Enhancement Incorporating Deep Neural Network
Tae Gyoong Kang\(^1\), Kisoo Kwon\(^1\), Jong Won Shin\(^2\), Nam Soo Kim\(^1\); \(^1\)Seoul National University, Korea; \(^2\)GIST, Korea
Thu-O-40-5

Recently, lots of algorithms using machine learning approaches have been proposed in the speech enhancement area. One of the most well-known approaches is the non-negative matrix factorization (NMF)-based one which analyzes noisy speech with speech and noise bases. However, NMF-based algorithms have difficulties in estimating speech and noise encoding vectors when their subspaces overlap. In this paper, we propose a novel speech enhancement algorithm which uses deep neural network (DNN) to improve the encoding vector estimation of the NMF-based technique. A DNN is trained to represent the mapping from noisy speech to corresponding encoding vectors. The quality of the enhanced speech from the proposed NMF-based scheme adopting DNN-based encoding vector estimation is compared with that from the conventional NMF-based technique. The experimental results showed that the proposed speech enhancement algorithm outperformed the conventional NMF-based speech enhancement technique.

A Data-Driven Approach to Speech Enhancement Using Gaussian Process
Sukanya Sonowal\(^1\), Kisoo Kwon\(^1\), Nam Soo Kim\(^1\), Jong Won Shin\(^2\); \(^1\)Seoul National University, Korea; \(^2\)GIST, Korea
Thu-O-40-6

This paper presents a novel data-driven approach to single channel speech enhancement employing Gaussian process (GP). Our approach is based on applying GP regression to estimate the residual gain with the input features being the a priori and a posteriori signal-to-noise ratios (SNRs). The residual gain is defined as the difference between the optimal gain and that obtained from the minimum mean-square error log-spectral amplitude (MMSE-LSA) estimator. Our proposed approach involves a cascaded structure consisting of two stages. At the first stage, the gain of the MMSE-LSA estimator is calculated in conjunction with the SNR features. In the second stage, the residual gains are estimated through GP and they are used to further enhance the output of the MMSE-LSA module. Experimental results show that the proposed approach produced better speech quality than not only the MMSE-LSA enhancement module but also the other data-driven technique.

**Notes**

- Room reverberation is a primary cause of failure in distant speech recognition (DSR) systems. In this study, we present a multichannel spectrum enhancement method for reverberant speech recognition, which is an extension of a single-channel dereverberation algorithm based on convolutive nonnegative matrix factorization (NMF). The generalization to a multichannel scenario is shown to be a special case of convolutive nonnegative tensor factorization (NTF). The presented algorithm integrates information from across different channels in the magnitude short time Fourier transform (STFT) domain. By doing so, it eliminates any limitations on the array geometry or a need for information concerning the source location, making the algorithm particularly suitable for distributed microphone arrays. Experiments are performed on speech data using actual room impulse responses from AIM database. Relative WER improvements using a clean-trained ASR system vary from +7.1\% to +30.1\% based on the number of channels and the source to microphone distances (1 to 3 meters).

- Speech Enhancement by Low-Rank and Convolutive Dictionary Spectrogram Decomposition
  - Zhao Chen, Brian McFee, Daniel P.W. Ellis; Columbia University, USA
  - Thu-O-40-3

- Multiple-Order Non-Negative Matrix Factorization for Speech Enhancement
  - Xabier Jaureguiberry\(^1\), Emmanuel Vincent\(^2\), Gaël Richard\(^1\); \(^1\)LTCI, France; \(^2\)INRIA, France
  - Thu-O-40-4

- NMF-Based Speech Enhancement Incorporating Deep Neural Network
  - Tae Gyoong Kang\(^1\), Kisoo Kwon\(^1\), Jong Won Shin\(^2\), Nam Soo Kim\(^1\); \(^1\)Seoul National University, Korea; \(^2\)GIST, Korea
  - Thu-O-40-5

- A Data-Driven Approach to Speech Enhancement Using Gaussian Process
  - Sukanya Sonowal\(^1\), Kisoo Kwon\(^1\), Nam Soo Kim\(^1\), Jong Won Shin\(^2\); \(^1\)Seoul National University, Korea; \(^2\)GIST, Korea
  - Thu-O-40-6

- This paper presents a novel data-driven approach to single channel speech enhancement employing Gaussian process (GP). Our approach is based on applying GP regression to estimate the residual gain with the input features being the a priori and a posteriori signal-to-noise ratios (SNRs). The residual gain is defined as the difference between the optimal gain and that obtained from the minimum mean-square error log-spectral amplitude (MMSE-LSA) estimator. Our proposed approach involves a cascaded structure consisting of two stages. At the first stage, the gain of the MMSE-LSA estimator is calculated in conjunction with the SNR features. In the second stage, the residual gains are estimated through GP and they are used to further enhance the output of the MMSE-LSA module. Experimental results show that the proposed approach produced better speech quality than not only the MMSE-LSA enhancement module but also the other data-driven technique.
Efficient Semi-Automatic Pronunciation Dictionary Bootstrapping

Tim Schlippe, Matthias Merz, Tanja Schultz; KIT, Germany

In this paper we propose efficient methods which contribute to a rapid and economic semi-automatic pronunciation dictionary development and evaluate them on English, German, Spanish, Vietnamese, Swahili, and Haitian Creole. First we determine optimal strategies for the word selection and the period for the grapheme-to-phoneme model retraining. In addition to the traditional concatenation of single phonemes most commonly associated with each grapheme, we show that web-derived pronunciations and cross-lingual grapheme-to-phoneme models can help to reduce the initial editing effort. Furthermore we show that our phoneme-level combination of the output of multiple grapheme-to-phoneme converters reduces the editing effort more than the best single converters. Totally, we report on average 15% relative editing effort reduction with our phoneme-level combination compared to conventional methods. An additional reduction of 6% relative was possible by including initial pronunciations from Wiktionary for English, German, and Spanish.

Rapidly Building Domain-Specific Entity-Centric Language Models Using Semantic Web Knowledge Sources

Murat Akbacak, Dilek Hakkani-Tür, Gokhan Tur; Microsoft, USA

For domain-specific speech recognition tasks, it is best if the statistical language model component is trained with text data that is content-wise and style-wise similar to the targeted domain for which the application is built. For state-of-the-art language modeling techniques that can be used in real-time within speech recognition engines during first-pass decoding (e.g., N-gram models), the above constraints have to be fulfilled in the training data. However collecting such data, even through crowd sourcing, is expensive and time consuming, and can still be not representative of how a much larger user population would interact with the recognition system. In this paper, we address this problem by employing several semantic web sources that already contain the domain-specific knowledge, such as query click logs and knowledge graphs. We build statistical language models that meet the requirements listed above for domain-specific recognition tasks where natural language is used and the user queries are about name entities in a specific domain. As a case study, in the movie domain where users’ voice queries are movie related, compared to a generic web language model, a language model trained with the above resources not only yields significant perplexity and word-error-rate improvements, but also presents an approach.
where such language models can be rapidly developed for other domains.

**Context-Dependent Pronunciation Error Pattern Discovery with Limited Annotations**

Ann Lee, James R. Glass; MIT, USA

A Computer-Assisted Pronunciation Training (CAPT) system can provide greater benefit to language learners if it provides not only scoring but also corrective feedback. However, the process of deriving pronunciation error patterns usually requires linguistic knowledge, or large quantities of expensive, annotated, corpora from nonnative speakers. In this paper we explore the possibility of deriving context-dependent error patterns with limited human annotations. A two-stage labeling mechanism is proposed, which first selects a set of templates for human annotation, and then propagates the labels. To deal with the imbalanced number of correct and incorrect phone-level pronunciations in nonnative speech, pronunciation patterns on an individual learner-level are first summarized, and then corpus-level clustering is done for template selection. The concept of contextual similarity based on a phonemic broad class definition is also proposed for label propagation. For evaluation, we view the task as an information retrieval task, and take advantage of metrics that consider both the importance and the ranking of an error type. Experimental results on a Chinese University of Hong Kong (CUHK) nonnative corpus show that the proposed framework can effectively discover prominent error patterns.

**Oral Session 42: Meta Data**

Peridet 206
13:30 - 15:30, Thursday 18 September 2014
Chairs: Khalid Chouki and Denis Jouvet

**Detecting Speaker Roles and Topic Changes in Multiparty Conversations Using Latent Topic Models**

Ashtosh Sapru, Hervé Bourlard; Idiap Research Institute, Switzerland

Accessing and browsing archives of multiparty conversations can be significantly facilitated by labeling them in terms of high level information. In this paper, we investigate automatic labeling of speaker roles and topic changes in professional meetings. Using the framework of unsupervised topic modeling we express speaker utterances as mixture of latent variables, each of which is governed by a multinomial distribution. The generated latent topic distributions are then used as features for predicting role and topic changes. Experiments performed on several hours of meeting data selected from AMI corpus reveal that latent topic features are effective predictors of speaker roles and topic changes. Furthermore, experiments also reveal an improvement in performance when latent topic information is combined with other multistream features.

**A Deep Neural Network Approach for Sentence Boundary Detection in Broadcast News**

Chenglin Xu 1, Lei Xie 1, Guangpu Huang 2, Xiong Xiao 2, Eng Siong Chng 2, Haizhou Li 2; 1Northwestern Polytechnical University, China; 2TL@NTU, Singapore

This paper presents a deep neural network (DNN) approach to sentence boundary detection in broadcast news. We extract prosodic and lexical features at each inter-word position in the transcripts and learn a sequential classifier to label these positions as either boundary or non-boundary. This work is realized by a hybrid DNN-CRF (conditional random field) architecture. The DNN accepts prosodic feature inputs and non-linearly maps them into boundary/non-boundary posterior probability outputs. Subsequently, the posterior probabilities are combined with lexical features and the integrated features are modeled by a linear-chain CRF. The CRF finally labels the inter-word positions as boundary or non-boundary by Viterbi decoding. Experiments show that, as compared with the state-of-the-art DT-CRF approach [1], the proposed DNN-CRF approach achieves 16.7% and 4.1% reduction in NIST boundary detection error in reference and speech recognition transcripts, respectively.

**Variable Span Disfluency Detection in ASR Transcripts**

Rahul Gupta 1, Sankaranarayanan Ananthakrishnan 2, Zhaojun Yang 1, Shrikanth S. Narayanan 1; 1University of Southern California, USA; 2Raytheon BBN Technologies, USA

Natural conversations often involve disfluencies in the form of revisions, repetitions, interjections, filled pauses and such. This paper focuses on word/phrase repetitions and revisions that are lexically well formed. These are generally captured by an ASR but pose problems to downstream processing such as spoken language translation (SLT). We describe a system to identify such word level disfluencies with a goal towards removing them in real time from the automatic recognition (ASR) system output. We use a span based training system to utilize the contextual information when tagging disfluencies. We design our system on the oracle transcripts and test them on both reference and ASR transcripts. We achieve an area under the receiver operating characteristics (ROC) curve for word level disfluency detection of .93 and .87 for the reference and the ASR transcripts respectively.

**A CRF-Based Approach to Automatic Disfluency Detection in a French Call-Centre Corpus**

Camille Dutrey 1, Chloé Clavel 2, Sophie Rosset 3, Ioana Vasilescu 3, Martine Adda-Decker 1; 1EDF, France; 2LTCI, France; 3LIMSI, France

In this paper, we present a Conditional Random Field based approach for automatic detection of edit disfluencies in a conversational telephone corpus in French. We define disfluency patterns into boundary/non-boundary posterior probability outputs. Subsequently, the posterior probabilities are combined with lexical features and the integrated features are modeled by a linear-chain CRF. The CRF finally labels the inter-word positions as boundary or non-boundary by Viterbi decoding. Experiments show that, as compared with the state-of-the-art DT-CRF approach [1], the proposed DNN-CRF approach achieves 16.7% and 4.1% reduction in NIST boundary detection error in reference and speech recognition transcripts, respectively.

**Multi-Pass Sentence-End Detection of Lecture Speech**

Madina Hasan, Rama Doddipatla, Thomas Hain; University of Sheffield, UK

Making speech recognition output readable is an important task. The first step here is automatic sentence end detection (SED). We
introduce novel F0 derivative-based features and sentence end distance features for SED that yield significant improvements in slot error rate (SER) in a multi-pass framework. Three different SED approaches are compared on a spoken lecture task: hidden event language models, boosting, and conditional random fields (CRFs). Experiments on reference transcripts show that CRF-based models give best results. Inclusion of pause duration features yields an improvement of 11.1% in SER. The addition of the F0-derivative features gives a further reduction of 3.0% absolute, and an additional 0.5% is gained by use of backward distance features. In the absence of audio, the use of backward features alone yields 2.2% absolute reduction in SER.

**Multi-Domain Disfluency and Repair Detection**

Victoria Zayats, Mari Ostendorf, Hannaneh Hajishirzi; University of Washington, USA

Thu-O-42-6

This paper investigates automatic detection of different types of self-repairs in spontaneous speech under different social contexts, from casual conversations to government hearings. The work shows that a simple CRF-based model is effective for cross-domain training, which is important for contexts where annotated data is not available. The approach explicitly represents common types of disfluencies observed in multi-domain data both in the model state space and the features extracted. In addition, the model incorporates an expanded state space for recognizing the repair structure, unlike prior work that annotates only the reparandum.

**Poster Session 28: Speech Synthesis III**

Max Atria Gallery
13:30 – 15:30, Thursday 18 September 2014
Chair: Zhenhua Ling

**Enabling Controllability for Continuous Expression Space**

Langzhou Chen, Norbert Braunischweiler; Toshiba Research Europe, UK

Thu-P-28-1, Poster

A continuous expression space assumes that each utterance contains individual expressions. Thus, it can be used to model detailed expression information in speech data. However, since an infinite number of different expressions can be contained in the continuous expression space, it is very difficult to manually label them. That means, these expressions are very hard to identify and to extract for synthesising expressive speech. A mechanism to control the continuous expression space is missing. In the discrete expression space though, only a few emotions are defined, thus users can easily choose from these emotions, but the range of expressivity is limited. This work proposes a method to automatically annotate expressions in the continuous expression space based on the cluster adaptive training (CAT) method. Using the proposed method, complex emotion information can be associated to the individual expressions in the continuous space. These emotion labels can be used as indexes of the expressions in the continuous space to enable users to select desired expressions at synthesis time, i.e. enable the controllability for the continuous expression space. Meanwhile, the rich expressive information in the continuous space is kept so that more expressive speech can be generated compared to the discrete space.

**Analysis of Spectral Enhancement Using Global Variance in HMM-Based Speech Synthesis**

Takashi Nose, Akinori Ito; Tohoku University, Japan

Thu-P-28-2, Poster

This paper analyzes the problem of the spectral enhancement technique using global variance (GV) in HMM-based speech synthesis. In the conventional GV-based parameter generation, spectral enhancement with variance compensation is achieved by considering a GV pdf with fixed parameters for every output utterances through the generation process. Although the spectral peaks of the generated trajectory are clearly emphasized and subjective clarity is improved, the use of the fixed GV parameters results in a much smaller variation of GVs among the synthesized utterances than that of the natural speech, which sometimes causes undesirable effect. In this paper, we examine the above problem in terms of multiple objective measures such as variance characteristics, spectral and GV distortions, and GV correlations and discuss the result. We propose a simple alternative technique based on an affine transformation that provides a closer GV distribution to the original speech and improves the correlation of GVs of generated parameter sequences. The experimental results show that the proposed spectral enhancement outperforms the conventional GV-based parameter generation in the objective measures.

**Intelligibility Analysis of Fast Synthesized Speech**

Cassia Valentini-Botinhao1, Markus Toman2, Michael Pucher2, Dietmar Schabus2, Junichi Yamagishi1; 1University of Edinburgh, UK; 2FTW, Austria

Thu-P-28-3, Poster

In this paper we analyse the effect of speech corpus and compression method on the intelligibility of synthesized speech at fast rates. We recorded English and German language voice talents at a normal and a fast speaking rate and trained an HMM-based synthesis system based on the normal and the fast data of each speaker. We compared three compression methods: scaling the variance of the state duration model, interpolating the duration models of the fast and the normal voices, and applying a linear compression method to generated speech. Word recognition results for the English voices show that generating speech at normal speaking rate and then applying linear compression resulted in the most intelligible speech at all tested rates. A similar result was found when evaluating the intelligibility of the natural speech corpus. For the German voices, interpolation was found to be better at moderate speaking rates but the linear method was even more successful at very high rates, for both blind and sighted participants. These results indicate that using fast speech data does not necessarily create more intelligible voices and that linear compression can more reliably provide higher intelligibility, particularly at higher rates.

**Speech Synthesis Reactive to Dynamic Noise Environmental Conditions**

Susana Palmaz López-Peláez, Robert A.J. Clark; University of Edinburgh, UK

Thu-P-28-4, Poster

This paper addresses the issue of generating synthetic speech in changing noise conditions. We will investigate the potential improvements that can be introduced by using a speech synthesiser that is able to modulate between a normal speech style and a speech style produced in a noisy environment according to a changing level of noise. We demonstrate that an adaptive system where the speech style is changed to suit the noise conditions maintains intelligibility and improves naturalness compared to traditional systems.
Partial Representations Improve the Prosody of Incremental Speech Synthesis
Timo Baumann; Universität Hamburg, Germany
Thu-P-28-5, Poster

When humans speak, they do not plan their full utterance in all detail before beginning to speak, nor do they speak piece-by-piece and ignoring their full message — instead humans use partial representations in which they fill in the missing parts as the utterance unfolds. Incremental speech synthesizers, in contrast, have not yet made use of partial representations and the information contained there-in.

We analyze the quality of prosodic parameter assignments (pitch and duration) generated from partial utterance specifications (substituting defaults for missing features) in order to determine the requirements that symbolic incremental prosody modelling should meet. We find that broader, higher-level information helps to improve prosody even if lower-level information about the near future is yet unavailable. Furthermore, we find that symbolic phrase-level or utterance-level information is most helpful towards the end of the phrase or utterance, respectively, that is, when this information is becoming available even in the incremental case. Thus, the negative impact of incremental processing can be minimized by using partial representations that are filled in incrementally.

Dialogue Context Sensitive Speech Synthesis Using Factorized Decision Trees
Pirros Tsiakoulis, Catherine Breslin, M. Gašić, Matthew Henderson, Dongho Kim, Steve Young; University of Cambridge, UK
Thu-P-28-9, Poster

This paper extends our recent work on rich context utilization for expressive speech synthesis in spoken dialogue systems in which significant improvements to the appropriateness of HMM-based synthetic voices were achieved by introducing dialogue context into the decision tree state clustering stage. Continuing in this direction, this paper investigates the performance of dialogue context-sensitive voices in different domains. The Context Adaptive Training with Factorized Decision trees (FD-CAT) approach was used to train a dialogue context-sensitive synthetic voice which was then compared to a baseline system using the standard decision tree approach. Preference-based listening tests were conducted for two different domains. The first domain concerned restaurant information and had significant coverage in the training data, while the second dealing with appointment bookings had minimal coverage in the training data. No significant preference was found for any of the voices when tested in the restaurant domain whereas in the appointment booking domain, listeners showed a statistically significant preference for the adaptively trained voice.

Concept-to-Speech Generation by Integrating Syntagmatic Features into HMM-Based Speech Synthesis
Xin Wang, Zhen-Hua Ling, Li-Rong Dai; USTC, China
Thu-P-28-7, Poster

In conventional concept-to-speech (CTS) methods, a common step is predicting abstract prosodic descriptions, such as the locations of accents and phrase boundaries, from the linguistic information provided by the text generation module. But the prediction results always contain errors, and unacceptable prosodic prediction may ruin the synthesized speech. In addition, linguistic information, which can be given conveniently and accurately by text generation, has not been directly utilized in the acoustic modelling and speech generation of CTS. This paper displays a CTS method utilizing HMM-based speech synthesis (HTS) and a text generation module called komet-Penn multilingual (KPML). In this method, syntagmatic features derived from the linguistic information given by KPML is directly added to the context features for context-dependent HMM modelling. Further, prosodic features are discarded during acoustic modelling to avoid costly prosodic annotation on the training waveforms and inaccurate prosodic prediction at synthesis time. Experiments show that the proposed method performs no worse than the conventional method with automatic prosodic prediction. When manual prosodic annotation on the training corpus is unavailable, the proposed method performs better.

On the Role of Missing Data Imputation and NMF Feature Enhancement in Building Synthetic Voices Using Reverberant Speech
Dhananjaya Govda1, Heikki Kallasjoki1, Reima Karhila1, Cristian Contan2, Kalle Palomäki1, Mircea Giurgiu2, Mikko Kurimo1; 1Aalto University, Finland; 2Universitatea Tehnică din Cluj-Napoca, Romania
Thu-P-28-8, Poster

In this paper, we study the role of a recently proposed feature enhancement technique in building HMM-based synthetic voices using reverberant speech data. The feature enhancement technique studied combines the advantages of missing data imputation and non-negative matrix factorization (NMF) based methods in cleaning up the reverberant features. Speaker adaptation of a clean average voice using noisy data is generally better than building a speaker dependent voice using the noisy data. In this paper, we show that the proposed feature enhancement technique can further improve the spectral match between the enhanced feature adapted voice and a clean speaker dependent voice.

Objective Evaluation of HMM-Based Speech Synthesis System Using Kullback-Leibler Divergence C.-T. Do1, M. Evrard1, A. Leman2, Christophe d’Alessandro1, Albert Rilliard1, J.-L. Crebouv2; 1LIMSI, France; 2Vocally, France
Thu-P-28-9, Poster

In this paper, we propose a new objective evaluation method for hidden Markov model (HMM)-based speech synthesis using Kullback-Leibler divergence (KLD). The KLD is used to measure the difference between the probability density functions (PDFs) of the acoustic feature vectors extracted from natural training and synthetic speech data. For the evaluation, Gaussian mixture model (GMM) is used to model the distribution of acoustic feature vectors, including the fundamental frequency (F0). Continuous F0, obtained with linear interpolation, is used in the evaluation. In essence, the KLD is the expectation of the logarithmic difference between the likelihoods calculated on training and synthetic speech. This likelihood difference is appropriate to characterize the quality of a HMM-based speech synthesis system in generating synthetic speech using a maximum likelihood criterion. The objective evaluation is tested with 3 different HMM-based speech synthesis systems which use multi-space distribution (MSD) to model discontinuous F0. These systems are trained on a common speech corpus in French. We propose an index to evaluate HMM-based speech synthesis system which takes into account the relative variation of the KLDs on test sets of synthetic and natural speech. This index correlates inversely with the result of the MOS (mean opinion score) perceptual test.

Notes

When manual prosodic annotation on the training corpus is unavailable, the proposed method performs better.

On the Role of Missing Data Imputation and NMF Feature Enhancement in Building Synthetic Voices Using Reverberant Speech
Dhananjaya Govda1, Heikki Kallasjoki1, Reima Karhila1, Cristian Contan2, Kalle Palomäki1, Mircea Giurgiu2, Mikko Kurimo1; 1Aalto University, Finland; 2Universitatea Tehnică din Cluj-Napoca, Romania
Thu-P-28-8, Poster

In this paper, we study the role of a recently proposed feature enhancement technique in building HMM-based synthetic voices using reverberant speech data. The feature enhancement technique studied combines the advantages of missing data imputation and non-negative matrix factorization (NMF) based methods in cleaning up the reverberant features. Speaker adaptation of a clean average voice using noisy data is generally better than building a speaker dependent voice using the noisy data. In this paper, we show that the proposed feature enhancement technique can further improve the spectral match between the enhanced feature adapted voice and a clean speaker dependent voice.

Objective Evaluation of HMM-Based Speech Synthesis System Using Kullback-Leibler Divergence C.-T. Do1, M. Evrard1, A. Leman2, Christophe d’Alessandro1, Albert Rilliard1, J.-L. Crebouv2; 1LIMSI, France; 2Vocally, France
Thu-P-28-9, Poster

In this paper, we propose a new objective evaluation method for hidden Markov model (HMM)-based speech synthesis using Kullback-Leibler divergence (KLD). The KLD is used to measure the difference between the probability density functions (PDFs) of the acoustic feature vectors extracted from natural training and synthetic speech data. For the evaluation, Gaussian mixture model (GMM) is used to model the distribution of acoustic feature vectors, including the fundamental frequency (F0). Continuous F0, obtained with linear interpolation, is used in the evaluation. In essence, the KLD is the expectation of the logarithmic difference between the likelihoods calculated on training and synthetic speech. This likelihood difference is appropriate to characterize the quality of a HMM-based speech synthesis system in generating synthetic speech using a maximum likelihood criterion. The objective evaluation is tested with 3 different HMM-based speech synthesis systems which use multi-space distribution (MSD) to model discontinuous F0. These systems are trained on a common speech corpus in French. We propose an index to evaluate HMM-based speech synthesis system which takes into account the relative variation of the KLDs on test sets of synthetic and natural speech. This index correlates inversely with the result of the MOS (mean opinion score) perceptual test.

Notes

When manual prosodic annotation on the training corpus is unavailable, the proposed method performs better.
The standard evaluation of intonation models is by means of non-referenced subjective tests (pair or MOS) in which subjects rate the quality or compare different samples without any explicit reference. These tests are usually conducted on an isolated sentence basis. However, for a single sentence, with no contextual information, there are multiple valid intonations. A subject’s preference over this range of intonation patterns may be highly personal. This paper investigates the degree to which this ambiguity in the appropriate intonation pattern impacts the assessments of prosody for speech synthesis systems. To examine this problem, the variance of the F0 pattern of several voiced sentences was modified and subjects asked to compare multiple versions with different levels of modification in terms of preference/quality. Then, they were presented with the reference which defines the original intonation and asked about the similarity to that reference. The results show that subjects can identify the samples with no F0 variance modification when given a reference but they don’t always prefer them. Thus, non-referenced tests with no context, though may help to analyse user acceptability, may not be appropriate to measure the performance of intonation models.

**Poster Session 29: Adaptation**

Max Atria Gallery  
13:30 - 15:30, Thursday 18 September 2014  
Chair: Wilson Tam

**Speaker Age Estimation for Elderly Speech Recognition in European Portuguese**

Thomas Pellegrini¹, Vahid Hedayati², Isabel Trancoso², Annika Hämläinen¹, Miguel Sales Dias³; ¹IRIT, France; ²INESC-ID Lisboa, Portugal; ³Microsoft, Portugal

Phone-like acoustic models (AMs) used in large-vocabulary automatic speech recognition (ASR) systems are usually trained with speech collected from young adult speakers. Using such models, ASR performance may decrease by about 10% absolute when transcribing elderly speech. Ageing is known to alter speech production in ways that require ASR systems to be adapted, in particular at the level of acoustic modeling. In this study, we investigated automatic age estimation in order to select age-specific adapted AMs. A large corpus of read speech from European Portuguese speakers aged 60 or over was used. Age estimation (AE) is based on i-vectors and support vector regression achieved mean error rates of about 4.2 and 4.5 years for males and females, respectively. Compared with a baseline ASR system with AMs trained using young adult speech and a WER of 13.9%, the selection of five-year-range adapted AMs, based on the estimated age of the speakers, led to a decrease in WER of about 9.3% relative (1.3% absolute). Comparable gains in ASR performance were observed when considering two larger age ranges (60-75 and 76-90) instead of six five-year ranges, suggesting that it would be sufficient to use the two large ranges only.

**Unsupervised Model Selection for Recognition of Regional Accented Speech**

Maryam Najafian¹, Andrea DeMarco², Stephen Cox², Martin Russell¹; ¹University of Birmingham, UK; ²University of East Anglia, UK

This paper is concerned with automatic speech recognition (ASR) for accented speech. Given a small amount of speech from a new speaker, is it better to apply speaker adaptation to the baseline, or to use accent identification (AID) to identify the speaker’s accent and select an accent-dependent acoustic model? Three accent-based model selection methods are investigated: the 'true' accent model, and unsupervised model selection using i-Vector and phonotactic-based AID. All three methods outperform the unadapted baseline. Most significantly, AID-based model selection using 43s of speech performs better than unsupervised speaker adaptation, even if the latter uses five times more adaptation data. Combining unsupervised AID-based model selection and speaker adaptation gives an average relative reduction in ASR error rate of up to 47%.

**Speaker Adaptation Based on Sparse and Low-Rank Eigenphone Matrix Estimation**

Wen-Lin Zhang¹, Dan Qu¹, Wei-Qiang Zhang², Bi-Cheng Li¹; ¹Zhengzhou Information Science & Technology Institute, China; ²Tsinghua University, China

The eigenphone based speaker adaptation outperforms the conventional MLLR and eigenvoice methods when the adaptation data is sufficient, but it suffers from severe over-fitting when the adaptation data is limited. In this paper, $l_1$ and nuclear norm regularization are applied simultaneously to obtain a more robust eigenphone estimation, resulting in a sparse and low-rank eigenphone matrix. The sparse constraint can reduce the number of free parameters while the low rank constraint can limit the dimension of phone variation subspace, which are both beneficial to the generalization ability. Experimental results show that the proposed method can improve the adaptation performance substantially, especially when the amount of adaptation data is limited.

**Multi-Accent Deep Neural Network Acoustic Model with Accent-Specific Top Layer Using the KLD-Regularized Model Adaptation**

Yan Huang, Dong Yu, Chaojun Liu, Yifan Gong; Microsoft, USA

We propose a multi-accent deep neural network acoustic model with an accent-specific top layer and shared bottom hidden layers. The accent-specific top layer is used to model the distinct accent specific patterns. The shared bottom hidden layers allow maximum knowledge sharing between the native and the accent models. This design is particularly attractive when considering deploying such a system to a live speech service due to its computational efficiency. We applied the KL-divergence (KLD) regularized model adaptation to train the accent-specific top layer. On the mobile short message dictation task (SMD), with 1K, 10K, and 100K British or Indian accent adaptation utterances, the proposed approach achieves 18.1%, 26.0%, and 28.5% or 16.1%, 25.4%, and 30.0% word error rate reduction (WERR) for the British and the Indian accent respectively against a baseline cross entropy (CE) model trained from 400 hour data. On the 100K utterance accent adaptation setup, comparable performance
gain can be obtained against a baseline CE model trained with 2000 hour data. We observe smaller yet significant WER reduction on a baseline model trained using the MMI sequence-level criterion.

A Low Complexity Model Adaptation Approach involving Sparse Coding Over Multiple Dictionaries

S. Shahnawazuddin, Rohit Sinha; IIT Guwahati, India

The work presented in this paper describes a novel on-line adaptation approach for extremely low adaptation data scenario. The proposed approach extends a similar redundant dictionary based approach reported recently in literature. In this work, the orthogonal matching pursuit (OMP) algorithm is used for bases selection instead of the matching pursuit (MP). This helps in avoiding the selection of an atom more than once. Furthermore, this work also explores the use of cluster-specific eigenvoices to capture local acoustic details unlike the conventional eigenvoices technique. These approaches are then combined to reduce the number of weight parameters being estimated for deriving adapted model. Towards this purpose, separate sparse coding of the test data is performed over a set of dictionaries. Those sparse coded supervectors are then scaled and used as the Gaussian mean parameter in the adapted model. Consequently, only a few scaling factors are needed to be estimated. Such a reduction in number of parameters is highly desirable for on-line applications where the latency is a major factor.

Effect of Frequency Weighting on MLP-Based Speaker Canonicalization

Yuichi Kubota, Motoi Omachi, Tetsuji Ogawa, Tetsunori Kobayashi, Tsuneo Nitta; Waseda University, Japan

Accurate and efficient speaker canonicalization is proposed to improve the performance of speaker-independent ASR systems. Vocal tract length normalization (VTLN) is often applied to speaker canonicalization in ASR; however, it requires parallel decoding of speech when estimating the optimal warping parameter. In addition, VTLN provides the same linear spectral transformation in an utterance, although optimal mapping functions differ among phonemes. In this study, we propose a novel speaker canonicalization using multi-layer perceptron (MLP) that is trained with a data set of vowels to map an input spectrum to the output spectrum of a standard speaker or a canonical speaker. The proposed speaker canonicalization operates according to the integration of MLP-based mapping and identity mapping that depends on frequency bands and achieves accurate recognition without any tuning of mapping function during run-time. Results of experiments conducted with a continuous digit recognition task showed that the proposed method reduces the intra-class variability in both of the vowel and consonant parts and outperforms VTLN.

Feature Space Maximum a posteriori Linear Regression for Adaptation of Deep Neural Networks

Zhen Huang1, Jinyu Li2, Sabato Marco Siniscalchi1, I-Fan Chen1, Chao Weng1, Chin-Hui Lee1; 1Georgia Institute of Technology, USA; 2Microsoft, USA

We propose a feature space maximum a posteriori (MAP) linear regression framework to adapt parameters for context dependent deep neural network hidden Markov models (CD-DNN-HMMs). Due to the huge amount of parameters used in DNN acoustic models in large vocabulary continuous speech recognition, the problem of over-fitting can be severe in DNN adaptation, thus often impair the robustness of the adapted DNN model. Linear input network (LIN) as a straight-forward feature space adaptation method for DNN, similar to feature space maximum likelihood linear regression (BM-LLR), can potentially suffer from the same robustness situation. The proposed adaptation framework is built based on MAP estimation of the LIN parameters by incorporating prior knowledge into the adaptation process. Experimental results on the Switchboard task show that against the speaker independent CD-DNN-HMM systems, LIN provides 4.28% relative word error rate reduction (WERR) and the proposed MAPLIN method is able to provide further 1.15% (totally 5.43%) WERR on top of LIN.

Speaker Adaptation of Context Dependent Deep Neural Networks Based on MAP-Adaptation and GMM-Derived Feature Processing

Natalia Tomashenko, Yuri Khokhlov; Speech Technology Center, Russia

In this paper we propose a novel speaker adaptation method for a context-dependent deep neural network HMM (CD-DNN-HMM) acoustic model. The approach is based on using GMM-derived features as the input to the DNN. The described technique of processing features for DNNs makes it possible to use GMM-HMM adaptation algorithms in the neural network framework. Adaptation to a new speaker can be simply performed by adapting an auxiliary GMM-HMM model used in calculation of GMM-derived features and can be regarded as adaptation in the feature space for a DNN system. In this work, traditional maximum a posteriori adaptation is performed for an auxiliary GMM-HMM model. Experiments show that the proposed adaptation technique can provide, on average, a 5%–36% relative word error reduction on different adaptation sets under supervised adaptation setup, compared to speaker independent (SI) CD-DNN-HMM systems. In addition, several multi-stream combination techniques are examined in order to improve the performance of the baseline SI model.

BUT 2014 Babel System: Analysis of Adaptation in NN Based Systems

Martin Karafiát, František Grézl, Karel Veselý, Mirko Hannemann, Igor Šzöke, Jan Černocký; Brno University of Technology, Czech Republic

Features based on a hierarchy of neural networks with compressive layers — Stacked Bottle-Neck (SBN) features — were recently shown to provide excellent performance in LVCSR systems. This paper summarizes several techniques investigated in our work towards Babel 2014 evaluations: (1) using several versions of fundamental frequency (F0) estimates, (2) semi-supervised training on un-transcribed data and mainly (3) adapting the NN structure at different levels. They are tested on three 2014 Babel languages with full GMM- and DNN-based systems. Separately and in combination, they are shown to outperform the baselines and confirm the usefulness of bottle-neck features in current ASR systems.

Speaker Adaptation of DNN-Based ASR with i-Vectors: Does it Actually Adapt Models to Speakers?

Mickael Rouvier, Benoit Favre; LIF (UMR 7279), France

Deep neural networks (DNN) are currently very successful for acoustic modeling in ASR systems. One of the main challenges with DNNs is unsupervised speaker adaptation from an initial speaker.
clustering, because DNNs have a very large number of parameters. Recently, a method has been proposed to adapt DNNs to speakers by combining speaker-specific information (in the form of i-vectors computed at the speaker-cluster level) with iMLLR-transformed acoustic features. In this paper we try to gain insight on what kind of adaptation is performed on DNNs when stacking i-vectors with acoustic features and what information exactly is carried by i-vectors. We observe on KERERE corpus that DNNs trained on i-vector features concatenated with iMLLR-transformed acoustic features lead to a gain of 0.7 points. The experiments show that using i-vector stacking in DNN acoustic models is not only performing speaker adaptation, but also adaptation to acoustic conditions.

Poster Session 30: Language Recognition
Max Atria Gallery
13:30 – 15:30, Thursday 18 September 2014
Chair: Yun Lei

Task-Aware Deep Bottleneck Features for Spoken Language Identification
Bing Jiang¹, Yan Song², Si Wei¹, Ian Vince McLaughlin¹, Li-Rong Dai¹; ¹USTC, China; ²Anhui USTC iFlytek, China
Thu-P-30-1, Poster

Recently, deep bottleneck features (DBF) extracted from a deep neural network (DNN) containing a narrow bottleneck layer, have been applied for language identification (LID), and yield significant performance improvement over state-of-the-art methods on NIST LRE 2009. However, the DNN is trained using a large corpus of specific language which is not directly related to the LID task. More recently, lattice based discriminative training methods for extracting more targeted DBF were proposed for ASR. Inspired by this, this paper proposes to tune the post-trained DNN parameters using an LID-specific training corpus, which may make the resulting DBF, termed a Discriminative DBF (D2BF), more discriminative and task-aware. Specifically, the maximum mutual information (MMI) criterion, with gradient descent, is applied to update the DNN parameters of the bottleneck layer in an iterative fashion. We evaluate the performance of the proposed D2BF using different back-end models, including GMM-MMI and i-vector, over the most confused 6 languages selected from NIST LRE 2009. The results show that the proposed D2BF is more appropriate and effective than the original DBF.

Virtual Example for Phonotactic Language Recognition
Rong Tong, Bin Ma, Haizhou Li; A*STAR, Singapore
Thu-P-30-2, Poster

One challenge in spoken language recognition is the availability of training data. In this paper, we propose a virtual example construction method to derive artificial training examples from the existing training data. Using the proposed method, both target virtual examples and non-target virtual examples can be derived from the available training samples. An iterative virtual example selection method is proposed to select those virtual examples that may provide extra discriminative information for language separation. By incorporating virtual examples in language classifier training, the language recognition performances are improved for both closed-set and open-set tasks. Specifically, for LRE 2009 evaluation data of three durations: 30-seconds, 10-seconds and 3-seconds, the language recognition performance improved by 3.67%, 11.98%, 6.42% respectively in closed-set conditions, and 10.14%, 10.55%, 5.75% respectively in open-set conditions.

Phonotactic Language Recognition Based on Time-Gap-Weighted Lattice Kernels
Wei-Wei Liu, Wei-Qiang Zhang, Jia Liu; Tsinghua University, China
Thu-P-30-3, Poster

Phonotactic method for spoken language recognition (SLR) deals with permissible phone patterns and their frequencies of occurrence in a specific language. Phone recognizers followed by vector space models (PR-VSM) system is a state-of-the-art phonotactic language identification system, in which any utterance can be mapped into a supervector filled with likelihood scores of the n-gram tokens (bag-of-n-gram). However, the bag-of-n-gram language model is not good at capturing the long-context co-occurrence relations due to the restriction match of the n-gram phonemes and vulnerable to the insert and delete errors induced by the frontend phone recognizer. We propose a novel approach to fill the gaps based on the use of time-gap-weighted lattice kernel (TGWLK) in this paper. The kernel is an inner product in the feature space generated by all contiguous and contiguous subsequences in variety length in the lattice, which are weighted by an exponentially decaying factor produced by the time gap length. The results of experiments on the NIST 2009 LRE corpus demonstrate that the proposed TGWLK shows a reduction in equal error rate (EER) than baseline system.

On the Complementarity of Short-Time Fourier Analyses of Different Lengths for Improved Language Recognition
Mireia Diez, Mikael Penaqarikano, German Bordel, Amparo Varona, Luis Javier Rodriguez-Fuentes; Universidad del País Vasco, Spain
Thu-P-30-5, Poster

Previous works have shown that remarkable performance improvements can be attained in speaker and language recognition tasks by combining several heterogeneous systems that provide complementary information. In this work, the complementarity of several i-vector language recognition systems, using Mel-Frequency Cepstral Coefficient (MFCC) features computed on Short-Time Fourier Analysis windows of different sizes, is studied. Language recognition experiments carried out on the NIST 2007 and 2009 LRE datasets reveal relative performance gains of up to 33%
when fusing the systems, with regard to the best single system. Results suggest that combining acoustic systems based on analysis windows of different sizes may allow to get advantage from both the sharper characterization of short events provided by short windows and the better frequency resolution of stationary events provided by long windows.

Modified-Prior i-Vector Estimation for Language Identification of Short Duration Utterances
Ruchir Travadi, Maarten Van Segbroeck, Shrikanth S. Narayanan; University of Southern California, USA

In this paper, we address the problem of Language Identification (LID) on short duration segments. Current state-of-the-art LID systems typically employ total variability i-Vector modeling for obtaining fixed length representation of utterances. However, when the utterances are short, only a small amount of data is available, and the estimated i-Vector representation will consequently exhibit significant variability, making the identification problem challenging. In this paper, we propose novel techniques to modify the standard normal prior distribution of the i-Vectors, to obtain a more discriminative i-Vector extraction given the small amount of available utterance data. Improved performance was observed by using the proposed i-Vector estimation techniques on short segments of the DARPA RATS corpora, with lengths as small as 3 seconds.

Language Recognition Using Phonotactic-Based Shifted Delta Coefficients and Multiple Phone Recognizers
Luis Fernando D’Haro, Ricardo Cordoba, Christian Salamea, Javier Ferreiros; Universidad Politécnica de Madrid, Spain

A new language recognition technique based on the application of the philosophy of the Shifted Delta Coefficients (SDC) to phone log-likelihood ratio features (PLLR) is described. The new methodology allows the incorporation of long-span phonetic information at a frame-by-frame level while dealing with the temporal length of each phone unit. The proposed features are used to train an i-vector based system and tested on the Albayzin LRE 2012 dataset. The results show a relative improvement of 33.3% in $C_{avg}$ in comparison with different state-of-the-art acoustic i-vector based systems. On the other hand, the integration of parallel phone ASR systems where each one is used to generate multiple PLLR coefficients which are stacked together and then projected into a reduced dimension are also presented. Finally, the paper shows how the incorporation of state information from the phone ASR contributes to provide additional improvements and how the fusion with the other acoustic and phonotactic systems provides an important improvement of 25.8% over the system presented during the competition.

PLLR Features in Language Recognition System for RATS
Oldřich Plchot1, Mireia Diez2, Mehdi Souffiar1, Lukáš Burget1; 1Brno University of Technology, Czech Republic; 2Universidad del País Vasco, Spain

In this paper, we study the use of features based on frame-by-frame phone posteriori (PLLRs) for language recognition. The results are reported on the datasets developed for the DARPA RATS (Robust Automatic Transcription of Speech) program, which seeks to advantage state of the art detection capabilities on audio from highly degraded communication channels. We show that systems based on the PLLRs outperform the standard acoustic system based on PLP2 features. By experimenting with the system combinations, we also demonstrate that the PLLR-based systems contain complementary information with respect to the PLP2 system. Finally we make a comparison between the PLLR and phonotactic systems with the outcome favorable to the PLLR.

Language Identification of Code Switching Sentences and Multilingual Sentences of Under-Resourced Languages by Using Multi Structural Word Information
Yin-Lai Yeong, Tien-Ping Tan; Universiti Sains Malaysia, Malaysia

Language identification (LID) is a process to identify the languages used in a text or speech. Code switching is the switching of a language in a sentence or speech utterance. This paper focuses on LID of words in code switching sentences. Code switching can occur intersentential or intrasential. The reasons why a writer switches from one language to another due to various reasons and among them are the inability to express opinion in a particular target language, to attract attention, to address different audience, habitual expressions and so on. The difficulty in identifying the languages of each word in a code switching sentence is because the languages have the same character set. In addition, code switching can happen in a sentence as short as a word or as long as a sentence. In this paper, we propose an automatic LID for words in code switching sentences by using multi structural word information (MUSWI) such as grapheme, syllable and word structure and calculate by using n-gram statistical model. The proposed MUSWI approach achieves 96.36% in term of accuracy on the code switching sentences, 99.07% on the multilingual sentences (non-code switching) which are under-resourced and closely related languages.

NOTES
### Author Index

<table>
<thead>
<tr>
<th>Author</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aalten, Pauline</td>
<td>106</td>
</tr>
<tr>
<td>Abbas, M.</td>
<td>108</td>
</tr>
<tr>
<td>Abdelaziz, Ahmed H.</td>
<td>107</td>
</tr>
<tr>
<td>Ahsudeh, Mohammad</td>
<td>108</td>
</tr>
<tr>
<td>Adda-Decker, Martine</td>
<td>120</td>
</tr>
<tr>
<td>Adel, Heike</td>
<td>118</td>
</tr>
<tr>
<td>Ahmed, Beena</td>
<td>127</td>
</tr>
<tr>
<td>Ahhara, Ryo</td>
<td>128</td>
</tr>
<tr>
<td>Ahirakinen, Manu</td>
<td>129</td>
</tr>
<tr>
<td>Akamine, Masami</td>
<td>142</td>
</tr>
<tr>
<td>Akarun, Lale</td>
<td>142</td>
</tr>
<tr>
<td>Akbacak, M.</td>
<td>133</td>
</tr>
<tr>
<td>Akiba, Tomoyosi</td>
<td>133</td>
</tr>
<tr>
<td>Akira, Hayakawa</td>
<td>148</td>
</tr>
<tr>
<td>Al, Jam</td>
<td>145</td>
</tr>
<tr>
<td>Albuquerque, Luciana</td>
<td>99</td>
</tr>
<tr>
<td>Al Dabel, Maryam</td>
<td>146</td>
</tr>
<tr>
<td>Alex, John</td>
<td>147</td>
</tr>
<tr>
<td>Al Hanai, Tuka</td>
<td>147</td>
</tr>
<tr>
<td>Al, Ahmed</td>
<td>147</td>
</tr>
<tr>
<td>Alispach, Samra</td>
<td>147</td>
</tr>
<tr>
<td>Al Kork, S.K.</td>
<td>147</td>
</tr>
<tr>
<td>Alku, Paavo</td>
<td>148</td>
</tr>
<tr>
<td>Almany, J.</td>
<td>150</td>
</tr>
<tr>
<td>Allauzen, Cyril</td>
<td>151</td>
</tr>
<tr>
<td>Aluisio, Sandra</td>
<td>151</td>
</tr>
<tr>
<td>Altumäe, Tanel</td>
<td>151</td>
</tr>
<tr>
<td>Alwan, Abeer</td>
<td>151</td>
</tr>
<tr>
<td>Ambikairajah, E.</td>
<td>151</td>
</tr>
<tr>
<td>Amelon, A.</td>
<td>151</td>
</tr>
<tr>
<td>Amino, Kanae</td>
<td>151</td>
</tr>
<tr>
<td>Amna, John</td>
<td>151</td>
</tr>
<tr>
<td>Amr, Toufik</td>
<td>151</td>
</tr>
<tr>
<td>A. Mistry, Hema</td>
<td>151</td>
</tr>
<tr>
<td>Ananthakrishnan, S.</td>
<td>151</td>
</tr>
<tr>
<td>Andersen, S.</td>
<td>151</td>
</tr>
<tr>
<td>André, B.</td>
<td>151</td>
</tr>
<tr>
<td>André-Obrecht, Régine</td>
<td>151</td>
</tr>
<tr>
<td>Andrews, Walter</td>
<td>151</td>
</tr>
<tr>
<td>Ane, G.</td>
<td>151</td>
</tr>
<tr>
<td>Anguera, Xavier</td>
<td>151</td>
</tr>
<tr>
<td>Anons, Jan-Niklas</td>
<td>151</td>
</tr>
<tr>
<td>Arai, Takayaki</td>
<td>151</td>
</tr>
<tr>
<td>Archanghi, Diana</td>
<td>151</td>
</tr>
<tr>
<td>Arias-Londoño, J.D.</td>
<td>151</td>
</tr>
<tr>
<td>Arik, Yasko</td>
<td>151</td>
</tr>
<tr>
<td>Arndt, Sebastian</td>
<td>151</td>
</tr>
<tr>
<td>Arenovitz, Hagai</td>
<td>151</td>
</tr>
<tr>
<td>Arsuikere, Harish</td>
<td>151</td>
</tr>
<tr>
<td>Artigas, M.</td>
<td>151</td>
</tr>
<tr>
<td>Astudillo, Ramon F.</td>
<td>151</td>
</tr>
<tr>
<td>Atkins, David C.</td>
<td>151</td>
</tr>
<tr>
<td>Attina, Virginia</td>
<td>151</td>
</tr>
<tr>
<td>Aubanel, Pauline</td>
<td>151</td>
</tr>
<tr>
<td>Aucher, Nadine</td>
<td>151</td>
</tr>
<tr>
<td>Auguste, Remi</td>
<td>151</td>
</tr>
<tr>
<td>Austel, Vernon</td>
<td>151</td>
</tr>
</tbody>
</table>

### B

<table>
<thead>
<tr>
<th>Author</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ab, Abhiram</td>
<td>134</td>
</tr>
<tr>
<td>Ab, Ramamo</td>
<td>134</td>
</tr>
<tr>
<td>Bachianer, Michel</td>
<td>134</td>
</tr>
<tr>
<td>Bach, Francis</td>
<td>134</td>
</tr>
<tr>
<td>Backström, Tom</td>
<td>134</td>
</tr>
<tr>
<td>Badin, Pierre</td>
<td>134</td>
</tr>
<tr>
<td>Bahaadini, Sara</td>
<td>134</td>
</tr>
<tr>
<td>Bai, Jie</td>
<td>134</td>
</tr>
<tr>
<td>Bailly, Gérard</td>
<td>134</td>
</tr>
<tr>
<td>Balasubramanian, A.</td>
<td>134</td>
</tr>
<tr>
<td>Ballard, Elaine</td>
<td>134</td>
</tr>
<tr>
<td>Ballard, Kipp</td>
<td>134</td>
</tr>
<tr>
<td>Banchs, Rafael E.</td>
<td>134</td>
</tr>
<tr>
<td>Banno, Hideki</td>
<td>134</td>
</tr>
<tr>
<td>Banzé, Desiré</td>
<td>134</td>
</tr>
<tr>
<td>Barbot, Nelly</td>
<td>134</td>
</tr>
<tr>
<td>Barker, J.</td>
<td>134</td>
</tr>
<tr>
<td>Barker, Tom</td>
<td>134</td>
</tr>
<tr>
<td>Barand, Einer</td>
<td>134</td>
</tr>
<tr>
<td>Barra-Chicote, Roberto</td>
<td>134</td>
</tr>
<tr>
<td>Bartels, Chris</td>
<td>134</td>
</tr>
<tr>
<td>Bates, Rebecca</td>
<td>134</td>
</tr>
<tr>
<td>Balintiner, Anton</td>
<td>134</td>
</tr>
<tr>
<td>Baumann, Timo</td>
<td>134</td>
</tr>
<tr>
<td>Bauer, Or Ak</td>
<td>134</td>
</tr>
<tr>
<td>Beat, Richard</td>
<td>134</td>
</tr>
<tr>
<td>Beaufays, Françoise</td>
<td>134</td>
</tr>
<tr>
<td>Beaufort, Richard</td>
<td>134</td>
</tr>
<tr>
<td>Bechet, Frederic</td>
<td>134</td>
</tr>
<tr>
<td>Bevan, Hamid</td>
<td>134</td>
</tr>
<tr>
<td>Bell, Peter</td>
<td>134</td>
</tr>
<tr>
<td>Bendris, Meriem</td>
<td>134</td>
</tr>
<tr>
<td>Bengio, Sanyo</td>
<td>134</td>
</tr>
<tr>
<td>Benin, Alberto</td>
<td>134</td>
</tr>
<tr>
<td>Bentz, Andre</td>
<td>134</td>
</tr>
<tr>
<td>Bernal, Brittany</td>
<td>134</td>
</tr>
<tr>
<td>Berry, James D.</td>
<td>134</td>
</tr>
<tr>
<td>Berry, Jeff</td>
<td>134</td>
</tr>
<tr>
<td>Bertrand, Thibaud</td>
<td>134</td>
</tr>
<tr>
<td>Beyer, André</td>
<td>134</td>
</tr>
<tr>
<td>Bialyi, Fadi</td>
<td>134</td>
</tr>
<tr>
<td>Bigot, Benjamin</td>
<td>134</td>
</tr>
<tr>
<td>Bill, Robert</td>
<td>134</td>
</tr>
<tr>
<td>Bimbot, Frederic</td>
<td>134</td>
</tr>
<tr>
<td>Birch, Alexander</td>
<td>134</td>
</tr>
<tr>
<td>Black, Matthew P.</td>
<td>134</td>
</tr>
<tr>
<td>Blavka, Karel</td>
<td>134</td>
</tr>
<tr>
<td>Blut, Thierry</td>
<td>134</td>
</tr>
<tr>
<td>Bocquet, Florent</td>
<td>134</td>
</tr>
<tr>
<td>Bohac, Marek</td>
<td>134</td>
</tr>
<tr>
<td>Bone, Daniel</td>
<td>134</td>
</tr>
<tr>
<td>Bonin, Francesca</td>
<td>134</td>
</tr>
<tr>
<td>Bordal, Guri</td>
<td>134</td>
</tr>
<tr>
<td>Bordel, German</td>
<td>134</td>
</tr>
<tr>
<td>Bor, Hynek</td>
<td>134</td>
</tr>
<tr>
<td>Bouafri, M.</td>
<td>134</td>
</tr>
<tr>
<td>Bouallegue, Mohammed</td>
<td>134</td>
</tr>
<tr>
<td>Bourlard, Hervé</td>
<td>134</td>
</tr>
</tbody>
</table>

### C

<table>
<thead>
<tr>
<th>Author</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cai, Lihong</td>
<td>136</td>
</tr>
<tr>
<td>Cai, Ming-Qi</td>
<td>136</td>
</tr>
<tr>
<td>Camelin, Nathalie</td>
<td>136</td>
</tr>
<tr>
<td>Campbell, Nick</td>
<td>136</td>
</tr>
<tr>
<td>Campbell, W.M.</td>
<td>136</td>
</tr>
<tr>
<td>Campion, Sébastien</td>
<td>136</td>
</tr>
<tr>
<td>Cancino Chacon, C.F.</td>
<td>136</td>
</tr>
<tr>
<td>Cano, Estefania</td>
<td>136</td>
</tr>
<tr>
<td>Cano, Xuan-Nga</td>
<td>136</td>
</tr>
<tr>
<td>Caraty, Marie-José</td>
<td>136</td>
</tr>
<tr>
<td>Cardinal, Patrick</td>
<td>136</td>
</tr>
<tr>
<td>Casado, Andrew</td>
<td>136</td>
</tr>
<tr>
<td>Castro Martínez, A.M.</td>
<td>136</td>
</tr>
<tr>
<td>Celikizkaya, Asli</td>
<td>136</td>
</tr>
<tr>
<td>Centelles, Jordi</td>
<td>136</td>
</tr>
<tr>
<td>Cernak, Milo</td>
<td>136</td>
</tr>
<tr>
<td>Černocký, Jan</td>
<td>136</td>
</tr>
<tr>
<td>Cerva, Petr</td>
<td>136</td>
</tr>
<tr>
<td>Čeh, Christian</td>
<td>136</td>
</tr>
<tr>
<td>Chi, Kian Ming A.</td>
<td>136</td>
</tr>
<tr>
<td>Chalopouka, Josep</td>
<td>136</td>
</tr>
<tr>
<td>Chab, Julian</td>
<td>136</td>
</tr>
<tr>
<td>Chang, John-Hyak</td>
<td>136</td>
</tr>
<tr>
<td>Chang, Shuangyung</td>
<td>136</td>
</tr>
<tr>
<td>Chang, Ying-Lan</td>
<td>136</td>
</tr>
<tr>
<td>Chappuis, Cyrielle</td>
<td>136</td>
</tr>
<tr>
<td>Charlet, Delphine</td>
<td>136</td>
</tr>
<tr>
<td>Cheung, Jian-Hong</td>
<td>136</td>
</tr>
<tr>
<td>Cho, Peter</td>
<td>136</td>
</tr>
<tr>
<td>Chou, Andrew</td>
<td>136</td>
</tr>
<tr>
<td>Chou, Andrew</td>
<td>136</td>
</tr>
<tr>
<td>Chou, Andrew</td>
<td>136</td>
</tr>
<tr>
<td>Chou, Andrew</td>
<td>136</td>
</tr>
</tbody>
</table>
Diamond Sponsors

NANYANG TECHNOLOGICAL UNIVERSITY
Temasek Laboratories@NTU

Tencent

Gold Sponsors

amazon.com
Google
Baidu
Microsoft
SinoVoice

Silver Sponsors

Unisound
DataTang
iFLYTEK
NUANCE
Vcyber.cn

Bronze Sponsors

ELRA
facebook
IBM Research
Raytheon
BBN Technologies
speechocean

Corporate Partner

SAMSUNG

Supporters

SINGAPORE EXHIBITION & CONVENTION BUREAU
Yandex
YourSingapore