Speaker Characterization using Average Filtering and Two Space Fusions

Chien-Lin Huang, Haizhou Li, Bin Ma
Institute for Infocomm Research, 1 Fusionopolis Way #21-01 Connexis (South Tower)
Singapore 138632
clhuang@i2r.a-star.edu.sg, hli@i2r.a-star.edu.sg, mabin@ i2r.a-star.edu.sg

Abstract

This paper presents average filtering and two space fusions approaches to speaker recognition. One of average filtering denotes the auto-regression moving average filtering for feature normalization. The other is a spectral average filter that averages short-time spectral features over a long-time window (SFLW) in an effort to capture the speaker traits that are manifested over a speech segment longer than a spectral frame. Two space fusions consider feature and model space fusions. In feature space, we study the fusion of multi-resolution feature extraction for an effective speaker characterization based on the same type of features. In model space, multi-order mixtures are used for the different amount of feature sample in speaker evaluation. The experiment conducted on the 2006 NIST Speaker Recognition Evaluation corpus shows that both average filtering and two space fusions achieve significantly EER reductions.

Keywords

speaker recognition; average filtering; two space fusion.

1 Introduction

Nowadays, voice biometrics has become increasing popular in telephony applications (Bimbot et al. 2004, Wu et al. 1997) The state-of-the-art text-independent speaker verification systems apply signal processing and statistical modeling techniques to characterize speakers. Speaker recognition task is typically formulated as a hypothesis test problem (NIST SRE 2009. Available: http://www.nist.gov/speech/tests/spk/).

There are three major components in a speaker recognition system: feature analysis, statistical modeling and verification decision. Short-time spectral features, such as Mel-frequency cepstral coefficients (MFCCs), are considered effective acoustic features for speaker recognition. Gaussian mixture models (GMM) (Reynolds et al. 2000) and support vector machine (SVM) (Campbell et al. 2006) shown the useful modeling techniques to reflect the statistics of sound patterns in speaker recognition. The basic classification decision is made by the log-likelihood ratio between the target speaker and a universal background model (UBM). In this paper, we are particularly interested in feature and model analysis for effective and efficiency speaker characterization.

In these frameworks, many efforts have been devoted to improving the effectiveness of MFCC which can be seen in several areas. One is to compensate the features to reduce the channel and noisy effect. The eigenchannel approach was studied (Kenny et al. 2007) where many different channel utterances of speakers were used to estimate speaker models and incorporate the channel information into speaker models. The other is more efficient to capture temporal features for speaker characteristics (Reynolds et al. 2003). The temporal features are manifested in a longer range of speech than a short-time segment. Moreover, feature mapping is used to compensate for channel mismatch (Reynolds et al. 2003). HLDA is applied to reduce the dimensionality of features and contains the discriminative power of features (Gales 1999). The extraction of frequency modulation (FM) components from a sub-band decomposition of speech signal was shown to be effective in speaker recognition (Nosratighods et al. 2009).

Score fusion is a popular technique in NIST SRE campaigns. We can take advantages of multiple heterogeneous subsystems for improved performance. The variety of feature and classification is applied for the compensation of difference techniques (Li et al. 2009, Kajarekar et al. 2009) For instance, the conventional MFCC, LPCC, PLP features and the GMM-UBM, kernel based SVM, GMM-SVM classification techniques are used for the overall performance improvement. The multiple kernel learning (MKL) is used to train a suitable weighting for fusing the output scores from multiple SVMs (Longworth et al. 2009).

Noise-robustness is one of important problems in speech signal processing. This paper furthers the feature studies by proposing a robust front-end analysis approach. We investigate a feature normalization processing consisting of mean subtraction, variance normalization, auto-regression moving average filtering (ARMA), and feature warping. ARMA is evidently good for noisy speech recognition (Chen et al. 2007). This study applies an auto-regression moving average filter in the cepstral domain.

Motivated by the findings in temporal features, we further the studies by proposing a spectral average filtering approach based on MFCC. The method of spectral average filtering is averaging short-time spectral features in a long-time window (SFLW). SFLW captures the spectral statistics over a long period of time. As a result, SFLW greatly reduces the amount of data involved in the modeling, testing, and thus the computational cost.

Moreover, this study proposed the scheme of two space fusion for speaker recognition including feature and model spaces. In feature space, the variety of feature sampling was used to consider the multi-resolution of speech signal. In model space, the multi-order mixtures were used to model the variety of feature sample. Based on the same dataset and the same type of feature, two space fusions are used to improve the performance of speaker recognition.

The rest of this paper is organized as follows. Section 2 introduces the average filtering for the front-end analysis of speaker recognition. Section 3 presents the scheme of two space fusions. The experimental results and analysis are presented in Section 4. Conclusions are finally drawn in Section 5.

2 Average Filtering

This study presents two kinds of average filters for the feature analysis of speaker recognition. One of average filtering denotes the auto-regression moving average filtering for the noise-robustness speech signal processing. The other is a spectral average filter that averages short-time spectral features over a long-time window (SFLW) in an effort to capture the speaker traits that are manifested over a speech segment longer than a spectral frame.

Each frame of the speech data is represented by a 36-dimensional feature vector, consisting of 12 MFCCs, along with their deltas, and double-deltas as the raw features. The feature normalization is applied to reduce the noise and channel effects after cepstral feature analysis. The first step of normalization processing is mean subtraction defined as

$$\bar{\mathbf{c}}_t = \mathbf{c}_t - \boldsymbol{\mu} , \quad \boldsymbol{\mu} = \sum_{t=1}^{T} \mathbf{c}_t / T$$

where $\boldsymbol{\mu}$ denotes a mean vector. This normalization is used to remove the global shift of the cepstral vectors. CMN compensates for the main effect of channel distortion and some of the side effects of additive noise (Torre et al. 2005). Moreover, variance normalization is estimated as follows

$$\hat{\mathbf{c}}_t = \bar{\mathbf{c}}_t / \boldsymbol{\sigma} , \quad \boldsymbol{\sigma} = \sqrt{\sum_{t=1}^{T} \left(\mathbf{c}_t - \boldsymbol{\mu}\right)^2 / T}$$

where $\boldsymbol{\sigma}$ is an estimate of the standard deviation. Additionally, the cepstral mean subtraction and cepstral variance normalization (MVN) are applied for slowly varying convolutional noises (Viikki et al. 1998).

After mean subtraction and variance normalization, the auto-regression moving average filtering is applied for further noisy reduction and defined by

$$\tilde{\mathbf{c}}_t = \left(\sum_{a=-A}^{t-1} \tilde{\mathbf{c}}_a + \sum_{a=t}^{A} \hat{\mathbf{c}}_a \right) / \left(2A+1\right)$$

where *A* is the order of the ARMA filter. Besides MVN, an auto-regression moving average filtering is directly applied in the cepstral domain. The ARMA filter is a low-pass filter, smoothing out any spikes in the time sequence (Chen et al. 2007).

Recently, the technology of feature warping shows a good improvement in speaker recognition (Burget et al. 2007, Huang et al. 2008). The goal of feature warping is to map the cepstral feature distribution over a specified speech interval so that the accumulated distribution is similar to a target distribution. Feature warping is used to reduce the additive noise and channel effects and to map a feature stream to a standard normal distribution. A lookup table is devised so as to map a rank order determined from the sorted cepstral feature elements to a warped feature using the desired warping target distribution (Pelecanos et al. 2001).

Figure 1 shows the visualization of cepstral domain plots of the time sequence of speech feature for the digit string “12345” in English.

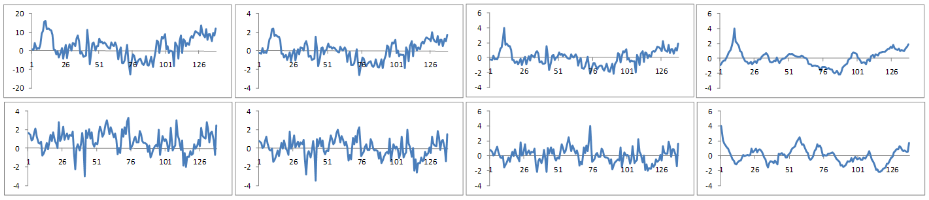


Figure 1. Cepstral domain plots of the time sequence of speech feature

The time sequence of the first row $\mathbf{c}^5$ and the second row $\mathbf{c}^{10}$ are plotted. The x-axis is time sequence and the y-axis means the log magnitude. Within each box, the first column shows original MFCC feature. MVN processing was shown in the second column. The third column shows MVW processing. Finally, the fourth column shows MVW processing with ARMA filtering. We can find a well-done noise reduction of MVW processing with ARMA filtering comparing with the original speech features. In theory, the small *A* will retain the short-time cepstral information but is more vulnerable to noise, while a large *A* will make the processed features less corrupted by noise but the short-time cepstral information will be lost. There is an inherent trade-off to decide the order *A* of the filter (Chen et al. 2007). We selected empirically in the experiments.

In feature analysis, speech signal is represented as a sequence of frames for short-time analysis. These frames are small enough to ensure the frequency characteristics of the magnitude spectrum are relatively stable. However, the sensation of a sound arises as the result of multiple short-time spectrums with different characteristics, such as vowel and consonant sections (Tzanetakis et al. 2002). To capture the spectral statistics over a long period of time, this paper proposes a way to average the short-time spectral features over a long-time window (SFLW). The SFLW features are estimated as the mean of multiple short-time spectral features over a long-time window. This transformation results in a more compact feature vector for statistical modeling as shown in Fig. 2.

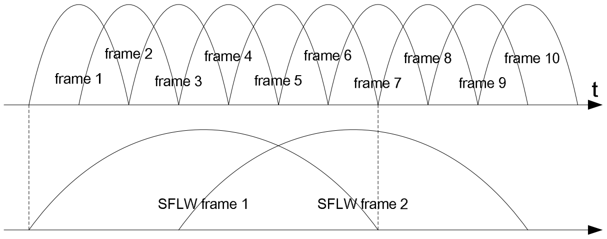


Figure 2. Illustration of feature analysis using short-time spectral features in a long-time window

With SFLW analysis, a sequence of feature vectors $X = \{x_1, x_2, ..., x_J\}$ is transformed to $V = \{v_1, v_2, ..., v_K\}$ where *J* and *K* denote the number of short-time frames and SFLW frames, respectively. Overlapping long-time windows are applied on top of the short-term features, reducing the *J* short-term frames to *K* SFLW frames, with $K = (J - L) / Z$. *L* denotes the size of the long-time window and *Z* is the step of the long-time window shift. The purpose of this transformation is to obtain a new representation of feature analysis which is more compact and suitable for statistical modeling. Due to the frame size reduction, SFLW allows much quicker speaker verification process, especially in statistical modeling and score normalization.

3 Two Space Fusions

Score fusion is a popular technique to take advantages of multiple heterogeneous subsystems for improved performance in (speaker recognition) NIST SRE campaigns. This study proposes the two space fusions including the fusion of multi-resolution in feature space and the fusion of multi-order mixtures in model space as shown in Fig. 3. We can consider this fusion as exploiting information across multi-resolution features and multi-order mixtures rather than different system architectures (Li et al. 2009).

This study proposes a fusion of multi-resolution acoustic features based on MFCC features using GMM-UBM system (Pelecanos et al. 2001). This study proposes a fusion of multi-resolutions acoustic features including the variety of feature sampling. The number of

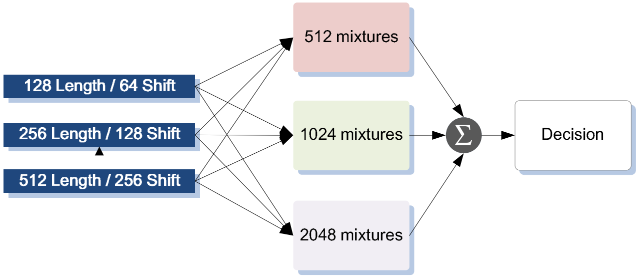


Figure 3. Illustration of two space fusions for speaker recognition

FFT sample points is usually a power of 2. Three resolutions are applied for the further fusion analysis in this study, including:

1) SFLW128: the short-time frequency analysis of 16 ms (128 samples at 8k Hz sampling rate and 64 sample shift) with long-time window of 48 ms containing 4 short-time frequency frames.

2) SFLW256: the short-time frequency analysis of 32 ms (256 samples at 8k Hz sampling rate and 128 sample shift) with long-time window of 96 ms containing 4 short-time frequency frames.

3) SFLW512: the short-time frequency analysis of 32 ms (512 samples at 8k Hz sampling rate and 256 sample shift) with long-time window of 192 ms containing 4 short-time frequency frames.

The goal of GMM is to match the distribution of observation data using multivariate Gaussians. Fusion of multi-order mixtures includes 512, 1024 and 2048 Gaussian models to fit the data distribution. We would like to see that informative speaker traits are available in spectral features of different frame rates. The fusion is a weighted sum of evaluation scores in different resolution features and multi-order mixtures. We set equal fusion weights for two space fusions. There is still a room for further improving the speaker recognition performance by linear fusion with prior-weighted Logistic Regression objective using FoCol tools (Available: http://www.dsp.sun.ac.za/~nbrummer/focal/index.htm).

4 Experiments and Results

Several concerning factors, such as language, various type of recording (telephone transmission type and microphone type) were examined in NIST SRE. The most collection is performed on telephone lines. All verification data adopted 8 kHz sampling frequency and 16 bit resolution for each sample. There are totally fifteen evaluation tasks in NIST SRE-2006 and noted with “TrainCondition-TestCondition” for the naming convention as shown in Table 1. In this study, we choose seven tasks for the experiments, including the core test which is the required test condition.

GMM-UBM classifier is used in this study which consists of a mixture of a specific number of multivariate Gaussian distributions. The NIST SRE-2004 one-side data was used to train the gender-dependent GMM-UBM models. The iterative EM algorithm is adopted to estimate the parameters of Gaussian components. A log-likelihood ratio (LLR) based evaluation function is applied for testing the trials as follows:

$$\Lambda = \frac{1}{T}\sum_{t=1}^{T}[\log p(v_t \mid \lambda_{SPK}) - \log p(v_t \mid \lambda_{UBM})]$$

		Test segment condition			
		10sec4w	1conv4w	1convmic	summed
Training condition	10sec4w	▨			
	1conv4w	▨	core test	✓	▨
	3conv4w	▨	✓	✓	▨
	8conv4w	✓	✓	✓	▨
	summed		▨		▨

Table 1. MATRIX OF TRAINING AND TEST SEGMENT CONDITIONS IN NIST SRE-2006

If the log-likelihood score is higher than the threshold $\Lambda > \theta$, the claimed speaker will be accepted, else rejected. The log-likelihood score is estimated by the Gaussian pdf as follows:

$$p(v_t \mid \lambda) = \sum_i^N w_i \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma_i|}} \exp[-\frac{1}{2}(v_t - u_i)^T \Sigma_i^{-1}(v_t - u_i)]$$

where v_t is the *t*-th test feature; u_i is the mean vector of the *i*-th Gaussian component; Σ_i represents the covariance matrix, and *d* denotes the dimension of the mean vector u_i. w_i is the mixture weight. In evaluation, only top one score is computed for each model and the score fusion is further applied. The eigenchannel approach was studied in (Kenny et al. 2007) in which a joint factor analysis model of intrinsic speaker variability and session variability in speaker recognition. The channel numbers was set as 30. Two types of eigenchannels were trained for telephone and microphone evaluation tasks.

4.1 Evaluation of the Order of ARMA Filter

This study evaluated the variety of windows to obtain a best result for ARMA filtering as shown in Fig. 5.

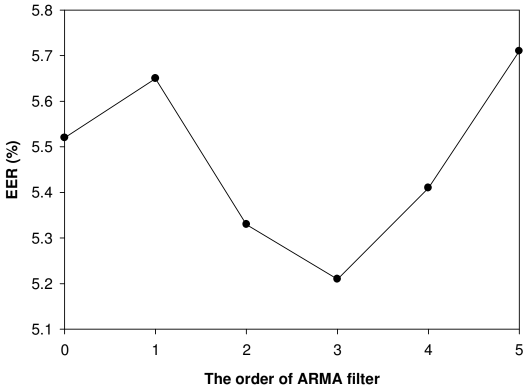


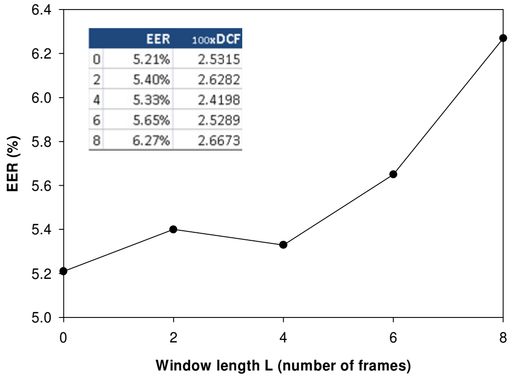
Figure 5. EER evaluations with the order of ARMA filter

The experiment was evaluated on the male part of NIST-SRE 2006 core test (1conv4w-1conv4w). The evaluation adopted 512 mixture numbers. The frame rate is 128 Length and 64 Shift. The case of $A = 0$ denotes no ARMA filtering and EER achieves

5.52%. The EER result was 5.21% on the best case of $A = 3$. Based on experiments, $A = 3$ was selected in this study. MVW processing was usually applied in the conventional feature extraction of speaker recognition. The visualization of spectral plots was shown in Fig 1. The proposed average filtering achieved the satisfied improvement for feature normalization. There was 5.62% relative EER reduction from 5.52% for without ARMA filtering to 5.21% for with ARMA filtering.

4.2 Number of the Spectral Average Filter

This study evaluated the variety of window lengths to obtain a best result in core test as shown in Fig. 6.



	EER	100×DCF
0	5.21%	2.5315
2	5.40%	2.6282
4	5.33%	2.4198
6	5.65%	2.5289
8	6.27%	2.6673

Figure 6. EER of core test evaluation with various window lengths

The experiment was evaluated on the male part of NIST-SRE 2006 core test. The evaluation adopted 512 mixture numbers. According that, the suitable long-time window was selected in the comparison of different window lengths. This study applies the short-time spectral analysis of 16 ms to obtain MFCC features, (128 samples at 8k Hz sampling rate and 64-sample shift) with half overlapping long-time window of 64 ms containing 4 short-time spectral frames to achieve the best performance. Although, the spectral average filter with 4 short-time average frames obtained 5.33% EER is slight worse than 5.21% EER without the spectral average filtering. The spectral average filter with 4 short-time frames achieves 2.4198 better than others conditions in $100 \times DCF$. The advantage of the spectral average filter is the feature sample reduction. As a result, the computation cost and storage are able to be more efficiency.

Figure 7 shows the statistical results of feature samples in 306 min speech data after VAD process. The number of feature samples was extracted from the amounts of the whole male GMM-UBM training data in NIST SRE-2004. This study compared the four conditions including Org123, SFLW128, SFLW256 and SFLW512. Org128 is applied as the baseline system. The sampling rate of Org128 is the short-time frequency analysis of 16 ms (128 samples at 8k Hz sampling rate and 64 sample shift) with long-time window of 48 ms without spectral average filtering. The number of feature sample of SFLW is 1,147,684 was translated to more than 50% feature sample reduction compared with that of Org128. Note that SFLW128 achieved a competitive performance with much reduced computation. The number of feature samples is greatly decreased with the larger frame size such as SFLW256 and SFLW512. The total frame size of the proposed fusion of multi-resolution is the sum of

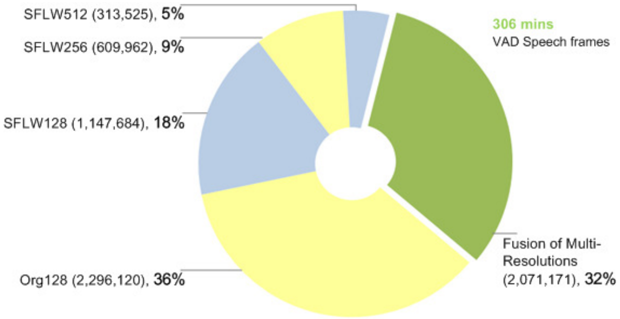


Figure 7. Number of feature samples for various types of features

frame size of SFLW128, SFLW256 and SFLW512. Moreover, the frame size of fusion of multi-resolutions is still smaller than the frame size of Org128.

4.3 Evaluation of Two Space Fusions

The frame rate in a typical speaker verification system is about 128 Length / 64 Shift (Org128) in the literature. This study used Org128 as the baseline, which has the largest number of feature samples, thus the highest computational cost. Table 2 shows the performance of NIST SRE-2006 evaluations at the different frame size and frame shift using the proposed average filtering and fusion of multi-resolution approach. The evaluation adopted 512 mixture numbers in the conditions of Org128, SFLW128, SFLW256, SFLW512 and Multi-Resolutions.

	Org128		SFLW128		SFLW256		SFLW512		Multi-Resolutions		Two Space Fusions	
	EER	100xDCF	EER	100xDCF	EER	100xDCF	EER	100xDCF	EER	100xDCF	EER	100xDCF
1conv4w-1conv4w	5.78%	2.8597	6.00%	2.7713	6.98%	3.0087	9.46%	4.0545	5.90%	2.6939	5.34%	2.5751
3conv4w-1conv4w	3.98%	2.0202	3.88%	2.0166	4.25%	2.1606	5.61%	2.7515	3.63%	1.9588	3.38%	1.9023
8conv4w-1conv4w	3.35%	1.5861	3.29%	1.4962	3.35%	1.4797	4.09%	1.9232	2.94%	1.3922	2.88%	1.3679
8conv4w-10sec4w	8.74%	3.8633	8.54%	3.9753	10.41%	4.5681	16.16%	6.5539	8.32%	3.6977	7.64%	3.4287
1conv4w-1convmic	7.47%	3.8940	7.48%	3.4051	8.80%	3.6901	11.35%	4.7156	7.02%	2.9510	6.20%	2.5722
3conv4w-1convmic	4.27%	2.3669	4.66%	2.4949	5.07%	2.6083	6.92%	3.1743	4.08%	1.9757	3.50%	1.8745
8conv4w-1convmic	4.48%	1.7403	4.02%	2.0045	4.48%	2.3031	6.42%	3.2932	3.33%	1.6169	3.39%	1.5928
Average	**5.44%**	**2.6186**	5.41%	2.5948	6.19%	2.8312	8.57%	3.7809	**5.03%**	**2.3266**	**4.62%**	**2.1876**

Table 2. Results of Different tasks on NIST SRE-2006

The experimental results indicate that the SFLW128 not only speeds up the processing but also improves the some of the system performance. SFLW consistently demonstrated favorable performance, while outperforming others in the conditions of "3conv4w-1conv4w", "8conv4w-1conv4w", "8conv4w-10sec4w" and "8conv4w-1convmic". In core test, the $100 \times DCF$ of SFLW128 is better than Org128.

The fusion of multi-resolutions can greatly reduce EER and DCF using the same type of features with different resolutions. Compared with Org128, the fusion of multi-resolutions showed the average 7.54% and 11.15% relative EER and reductions, respectively. The results show that the multi-resolution features are complementary. Therefore, the fusion of multi-resolutions computation efficient and achieves lower DCF and EER to speaker recognition. In Table 2, the last column showed the two space fusions. The results showed that the average 15.07 relative EER reduction from 5.44% EER for Org128 to 4.62% EER for two space fusions. There was average 16.46 relative $100\times DCF$ reduction from 2.6186 for Org128 to 2.1876 for two space fusions. We set equal fusion weights for two space fusions.

5 Conclusion

This study presented novel approaches for robust speaker recognition including average filtering and two space fusions. The average filtering means the ARMA filtering for feature normalization and a spectral average filter (SFLW). We propose a feature analysis strategy that averages short-time spectral features over a long-time window (SFLW) in an effort to capture the speaker traits that are manifested over a speech segment longer than a spectral frame. SFLW achieves more efficient speaker verification and the computation cost is greatly reduced. There is more than half feature sample reduction compared to the commonly used feature sampling (128 Length and 64 Shift). There was 5.62% relative EER reduction from 5.52% for without ARMA filtering to 5.21% for with ARMA filtering. Two space fusions denote the fusion of multi-resolutions in feature space and the fusion of multi-order mixtures in model space. Fusion of multi-resolutions provides a good way to improve a single system trained with the same type of features using different frame rates. In NIST SRE-2006, the two space fusions provide a further improvement resulting in average 15.07% and 16.46% relative EER and DCF reductions, respectively.

6 References

Bimbot, F.; Bonastre, J.-F.; Fredouille, C.; Gravier, G.; Magrin-Chagnolleau, I.; Meignier S.; Merlin T.; Ortega-Garcia, J.; Petrovsk-Delacrtaz, D. and Reynolds, D., “A tutorial on text-independent speaker verification,” EURASIP J. Appl. Signal Processing, vol. 4, pp. 430–451, 2004.

Wu C.-H. and Chen J.-H., “Speech activated telephony e-mail reader (SATER) based on speaker verification and text-to-speech conversion,” IEEE Transactions on Consumer Electronics, vol.43, no.3, pp.707–716, 1997.

Reynolds D. A.; Quatieri T. F. and Dunn R. B., “Speaker verification using adapted Gaussian mixture modeling,” Digital Signal Processing, vol. 10, pp. 19–41, 2000.

Campbell W. M.; Campbell J. P.; Reynolds D. A.; Singer E. and Torres-Carrasquillo P. A., “Support vector machines for speaker and language recognition,” Computer Speech and Language, vol. 20, pp. 210–229, 2006.

Kenny P.; Boulianne G.; Ouellet P. and Dumouchel P., “Joint factor analysis versus eigenchannels in speaker recognition,” IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 4, pp.1435–1447, 2007.

Reynolds D.; Campbell W.; Campbell J.; Dunn B.; Gleason T.; Jones D.; Quatieri T.; Quillen C.; Sturim D. and Torres-Carrasquillo P., “Beyond cepstra: exploiting high-level information in speaker recognition,” Workshop on Multimodal User Authentication, 2003.

Reynolds D. A., “Channel robust speaker verification via feature mapping,” in Proc. ICASSP, pp. 53–56, Hong Kong, 2003.

Gales M.J.F., “Semi-tied covariance matrices for hidden Markov models,” IEEE Transactions on Speech Audio Processing, vol. 7, pp. 272–281, 1999.

Nosratighods M.; Thiruvaran T.; Epps J.; Ambikairajah E.; Ma B. and Li H., “Evaluation of a Fused FM and Cepstral-Based Speaker Recognition System on the NIST 2008 SRE,” in Proc. ICASSP, pp. 4233–4236, Taipei, Taiwan, 2009.

Li H.; Ma B.; Lee K.-A.; Sun H.; Zhu D.; Sim K. C.; You C.; Tong R.; Karkkainen I.; Huang C.-L.; Pervouchine V.; Guo W.; Li Y.; Dai L.; Nosratighods M.; Tharmarajah T.; Epps J.; Ambikairajah E.; Chng E.-S.; Schultz T. and Jin Q., “The I4U System in NIST 2008 Speaker Recognition Evaluation,” in Proc. ICASSP, pp. 4201–4204, Taipei, Taiwan, 2009.

Kajarekar S. S.; Scheffer N.; Graciarena M.; Shriberg E.; Stolcke A.; Ferrer L. and Bocklet T., “The SRI NIST 2008 Speaker Recognition Evaluation System,” in Proc. ICASSP, pp. 4205–4208, Taipei, Taiwan, 2009.

Longworth C. and Gales M.J.F., “Combining Derivative and Parametric Kernels for Speaker Verification,” IEEE Transactions on Audio, Speech and Language Processing, vol. 17, no. 4, pp. 748–757, 2009.

Chen C.-P. and Bilmes J., "MVA Processing of Speech Features," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 1, pp. 257–270, 2007.

Torre A.; Peinado A. M.; Segura J. C.; Perez-Cordoba J. L.; Bentez M. C. and Rubio A. J., “Histogram Equalization of Speech Representation for Robust Speech Recognition,” IEEE Transactions on Speech and Audio Processing, vol. 13, no. 3, pp. 355–366, 2005.

Viikki O. and Laurila K., “Cepstral domain segmental feature vector normalization for noise robust speech recognition,” Speech Communication, vol. 25, pp. 133–147, 1998.

Burget L.; Matejka P.; Schwarz P.; Glembek O. and Cernocký J., “Analysis of Feature Extraction and Channel Compensation in a GMM Speaker Recognition System,” IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 7, pp. 1979–1986, 2007.

Huang C.-L.; Ma B., Wu C.-H.; Mak B. and Li H., “Robust Speaker Verification Using Short-Time Frequency with Long-Time Window and Fusion of Multi-Resolutions,” in Proc. Interspeech, pp. 1897–1900, Brisbane, Australia, 2008.

Pelecanos J. and Sridharan S., “Feature warping for robust speaker verification,” in Proc. 2001: A Speaker Odyssey, pp. 213–218, 2001.

Tzanetakis G. and Cook P., “Musical genre classification of audio signals,” IEEE Transactions on Speech and Audio Processing, vol. 10, no. 5, pp. 293–302, 2002.