Effectiveness of using language independent transcribers for spoken language identification for different Indian languages

Rajlakshmi Saikia, Sanasam Ranbir Singh and Priyankoo Sarmah
Indian Institute of Technology Guwahati
saikia.rajlakshmi349@gmail.com, ranbir@iitg.ernet.in, priyankoo@iitg.ernet.in

Abstract

Automatic spoken language identification systems are usually built by extracting suitable acoustic features from speech samples, and in some cases by building a language dependent speech-to-text. Acoustic features are highly vulnerable to recording environments and recording channels, often affecting performance. On the other hand, building language dependent speech-to-text synthesizer is a resource intensive and expensive task. Instead of using language dependent speech-to-text transcribers for language identification, this paper investigates the efficacy of using language independent speech-to-text transcribers (trained on languages different from the target languages) in automatic language identification for various Indian languages (namely Assamese, Bengali, Hindi, Gujarati, Manipuri, Mizo, Tamil and Telegu) deploying various text classification methods. It further compares the performances obtained over text generated by language independent transcribers with that of the classifiers built using audio features and manually transcribed text. From various experimental setups over speech samples recorded under controlled (studio recorded) and uncontrolled environment (outdoor recording), it is evident that transcription text using language independent transcriber can effectively be used for spoken language identification tasks. .

Keywords

Automatic language identification, language independent transcriber, speech-to-text synthesizer, Indian languages.

1. Introduction

The task of automatic spoken language identification (LID) in speech is to identify the language being spoken from speech samples by unknown speakers (Jothilakshmi 2010). Existing studies in spoken LID systems have resorted to two different directions; (i) using various speech features such as Linear Prediction Cepstral Coefficients (LPCCs) (Cimarusti and Ives 1982), pitch and energy contours (Cummins et al. 1999), Mel-Frequency Cepstral Coefficients (MFCCs) (Li 1994) etc., and (ii) using transcribed text either through phonetic transcription (House and Neuburg 1977) or language specific ASR (speech-to-text synthesizers) (Safitri et al. 2016). Study in (Hughes et al. 2006) presents the fact that LID with audio features is a complex task considering the large variations in speaker, speaking style and tone. Another important challenge in LID with audio feature is that it is highly vulnerable to speaker, back ground noise and environment. Though LID through transcribed text using ASR is claimed

to be easier, building ASR for target languages is an expensive task and ASR for many resource poor Indian languages are not available. In this study, we deploy language independent transcribers (LIT) over speech samples collected for various Indian languages and investigate their performances on LID problem. This paper is the extended study of our earlier worked published in (Saikia et al. 2017). A LIT system is a speech-to-text synthesizer which is built over a known source language and is used to transcribe speech samples of different languages. An illustrative scenario of a language independent transcriber is described below.

We request a random person to transcribe a speech spoken in a language that he/she has never heard before (without revealing the language identity). We request him/her to transcribe the sounds he/she perceives in a script he/she can read. Transcribed text may not have any similarity with that of the correct transcription.

While obtaining text from language independent transcribers, we are expecting following characteristics inherited to the transcribed text:

- Homogeneous transcription across heterogeneous speech samples, by virtue of using a common speech-to-text transcriber.
- High coherent intra-language transcription error and low coherent inter-language transcription error.
- Discriminant inter-language transcription error characteristics capturing language information.

After obtaining homogeneous transcribed text for all the speech samples in different languages using a LIT, the text samples are subjected to various classification frameworks (details in section 3). From various experimental observations over both studio recording and outdoor recording data sets, it is observed that the transcription text obtained from LIT can be effectively used for LID tasks and provide comparable performances (even better in many instances) with that of audio features.

Rest of the paper is organized as follows. Some of the previous related works are discussed in Section 2. Section 3 and Section 4 describe the proposed frameworks and experimental data sets respectively. Language discriminative characteristics of the transcribed text is analyzed in section 5. Experimental results are discussed in Section 6. Paper concludes in Section 7.

2. Related Work

In literature, studies related to LID have considered wide ranges of methods and approaches. Some of the widely used approaches consider features such as pitch contours, formant vectors and other acoustics and prosodic features (Thymé-Gobbel and Hutchins 1996), (Hazen and Zue 1997). In (Yan and Barnard 1995), authors consider three types of models - acoustic model, language model and duration model to capture the acoustic, phonotactic and prosodic information. There have been few studies for Indian language identification tasks. Vector Quantization based methods are used to model 5 (five) Indian languages namely Tamil, Telegu, Malayalam, Kannada and Hindi (Balleda et al. 2000) with an average identification accuracy of 84% for all 5 Indian languages. Further, autoassociative neural network (ANN) is explored for capturing language specific features for four Indian languages namely Tamil, Telegu, Kannada and Hindi (Mary and Yegnanarayana 2004). Considering spectral features, authors reported an identification accuracy of 100% for Hindi and Telegu. In (Jothilakshmi et al. 2012), authors adopted hierarchical approach for identifying 9 (nine) Indian languages. It first identifies language group and then the target language within the group by considering MFCC, MFCC with delta and acceleration coefficient and shifted delta cepstral.

Unlike above studies, few studies have investigated the effect of using transcription text. In (House and Neuburg 1977), authors use manually transcribed text for eight languages and build HMM based LID system. Instead of using acoustic features extracted from speech sig-

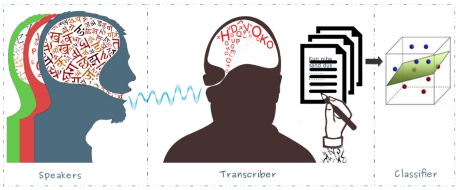

Figure 1: Transcript generation process

nals, information related to languages are extracted from broad phonetic transcription. Speech signal is considered as a sequence of symbols from top consonant, fricative consonant, vowel and silence. Observations in this study has shown capability of performing LID task using transcription text. Authors in (Safitri et al. 2016) adopted phonotactic approach to identify language from spoken data. To map the speech signals to the corresponding language, authors have created statistical phonotactic models. This study has been done on three local Indonesian languages namely: Mi-nangkabau, Sundanese and Javanese using Phone Recognition followed by Language Modelling (PRLM) and Parallel Phone Recognition followed by Language Modelling (PPRLM) methods. Authors reported that PRLM method is able to provide best accuracy while using the phone recognizer which is trained on English and Russian speech data with the average of 77.42% and 75.94% respectively.

LID from spoken speech is a well explored area and majority of the studies consider audio features. Very less studies focused on LID considering textual context of the speech samples. Authors in (Balleda et al. 2000) and (Safitri et al. 2016) consider textual features for LID by considering manually transcribed phonetic data and building language dependent phone to speech recognizer respectively. Unlike these studies, our proposed study uses unbiased or language independent transcriber which has no knowledge of the target languages to classify.

3. Proposed Framework

Figure 1 shows thematic representation of the proposed language independent transcriber to generate homogeneous text across heterogeneous speech samples. The speaker depicts different heterogeneous speech samples (different languages, different speakers). The transcriber depicts an unbiased transcriber (independent of the incoming language) who has no knowledge of the speaker and language of the incoming speech samples. The transcriber generates homogeneous text patterns representing the sound he/she perceives in a script he/she understands. The classification models are then built over the transcribed text. To simulate the unbiased transcriber, this study explores three standard speech-to-text transcribers which are explained below:

1. **IBM Watson Speech to Text API** [1]: The IBM Watson Speech to Text service uses speech recognition capabilities to convert Arabic, English, Spanish, French, Brazilian Portuguese, Japanese, and Mandarin speech into text. It transcribes speech from various languages and audio formats to text with low latency. For most languages, the service supports two sampling rates, broadband and narrow band. Some of the output features of this synthesizer are keyword spotting, word alternatives, confidence, and time stamps, maximum alternatives and interim results. To simulate homogeneous transcription, this study considers English transcription from the non-English inputs (Indian languages).

[1] https://www.ibm.com/watson/services/speech-to-text/

Table 1: Examples of text generated using transcriber for a speech sample of Assamese language

Correct Transcript	Transcriber Generated Transcript	Transcriber
Photo bur dekhuwai thokar majote kobo loi dhorile	Photo booth required for kamate hot overload really	Web Speech API
Photo bur dekhuwai thokar majote kobo loi dhorile	For two more days quite took on model to it hard global really	IBM Watson
Photo bur dekhuwai thokar majote kobo loi dhorile	hh r g hh r hh hh r hh k t r hh d g t m hh g r g r uw r	HMM based

2. **Web Speech API**[2]: This API recognizes over 80 languages and variants. It has two parts: SpeechSynthesis (Text-to-Speech), and SpeechRecognition (Asynchronous Speech Recognition). Like in IBM Watson, to generate homogeneous transcription text we consider Asynchronous Speech Recognition API in English language.
3. **HMM based synthesizer**: This speech-to-text synthesizer is built using Hidden Markov Model over Indian English in the EMST Lab at IIT Guwahati [3]. The synthesizer is built using HTK [4] tool over a NDTV news (speech) data set locally generated at EMST Lab. It generates phone level text without placing word boundaries.

Table 1 presents an example of text generated using the above transcribers over Assamese speech samples. From the table, it is observed that, using the text generated by these transcribers are totally different from the actual (correct) text. Only Web Speech API is able to produce transcript having one word common with the correct text.

3.1. Classification Frameworks

Every speech sample s_i in the training data set is represented as $< \bar{x}_i, y_i >$, where $\bar{x}_i$ is the feature vector defined over unique unigrams in the transcription corpus, and y_i is the language label. Similarly, test speech sample s_j is transcribed using the same transcriber used for generating training corpus and is defined as $< \bar{x}_j, ? >$, where $\bar{x}_j$ is feature vector defined over unigram features of the testing corpus. We consider popular classifiers namely k-Nearest Neighbour (KNN) (Tam et al. 2002), Random Forest (RF) (Khan et al. 2010), Decision Tree (DT), Support Vector Machine with linear kernel(SVM) (Vapnik 2013), Naive Bayes (NB), (Rish et al. 2001), and Recurrent Neural Network (RNN) (Sutskever 2013). NB with Multinomial kernel is denoted by NB1, and that of Gaussian kernel is denoted by NB2.

3.2. Decision Fusion Based Ensembling Framework

In addition to the individual classifiers mentioned above, we further adapt a decision fusion based ensembling (DFBE) framework, as illustrated in Figure 2, to investigate the combined effects of several classifiers and transcribers. Similar framework has also been explored in multi-model classification studies (Skowron et al. 2006). As shown in the Figure 2, the ensembling framework uses two levels of classifiers. The task of the classifiers in the first level is to provide decision on the language from the transcribed text. The classifier at the second level fuses the decision of various classifiers in the first level to provide better judgment on identifying the language.

[2]https://www.google.com/chrome/demos/speech.html
[3]https://www.iitg.ac.in/eee/emstlab/index.php
[4]http://htk.eng.cam.ac.uk/

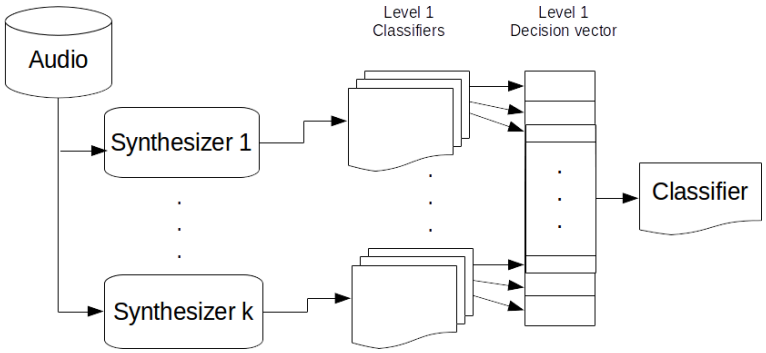

Figure 2: Ensemble framework

Table 2: Description of the Speech Data set

Language	Studio Recorded			YouTube		
	Dur	#Sent	#Speaker	Dur	#Sent	#Speaker
Assamese	3 hrs	2890	1	6 hrs	4000	10
Manipuri	3 hrs	3235	1	6 hrs	3000	6
Bengali	3 hrs	3179	1	6 hrs	3000	8
Hindi	3 hrs	2302	1	6 hrs	5863	10
Tamil	3 hrs	3600	1	6 hrs	5879	10
Telegu	3 hrs	2481	1	6 hrs	5418	10
Gujarati	3 hrs	3500	1	6 hrs	5089	10
Mizo	2 hrs	2800	1	6 hrs	-	-

4. Datasets

The experimental results reported in this study consider two type of datasets with different characteristics as described below.

Controlled studio recorded data: We have mainly considered studio recorded data for eight (8) Indian languages namely, Assamese, Manipuri, Bengali, Hindi, Tamil, Telegu, Gujarati and Mizo. All these samples are recorded in studio under controlled environment. For one language, one male speaker is considered. Technical settings for the recording are 16 bits mono channel with sampling rate 48000 Hz. Recorded audio samples are then converted to 16000 Hz. Table 2 presents the characteristics of this data set.

Youtube data: We further considered speech samples collected from Youtube for the same seven (7) Indian languages except Mizo. Samples collected from Youtube are group discussion with multiple speakers, and have background noises like sound, music etc. These speech samples are in 44100 Hz and stereo channel. Before using these for LID, we convert them to 16000 Hz and mono channel. Each raw speech Youtube file is of 50 min to 80 min duration. Each speech samples are segmented to multiple smaller length speech samples. For our experiments, speech samples of two different segment lengths (10sec and 30sec) are considered. Table 2 presents the statistics of this data set.

5. Text Analysis

The idea of using language independent speech-to-text transcriber is that though the transcriber may not generate the correct transcription of the input speech samples but it will generate a homogeneous text sequence as it perceives. The sequence of text pattern that

Table 3: Different Pairs of Language Combinations

L1	Assamese + Bengali	L8	Bengali + Hindi	L15	Manipuri + Telegu
L2	Assamese + Manipuri	L9	Bengali + Gujarati	L16	Gujarati + Hindi
L3	Assamese + Hindi	L10	Bengali + Tamil	L17	Gujarati + Tamil
L4	Assamese + Gujarati	L11	Bengali + Telegu	L18	Gujarati + Telegu
L5	Assamese + Tamil	L12	Manipuri + Hindi	L19	Tamil + Hindi
L6	Assamese + Telegu	L13	Manipuri + Gujarati	L20	Tamil + Telegu
L7	Bengali + Manipuri	L14	Manipuri + Tamil	L21	Telegu + Hindi

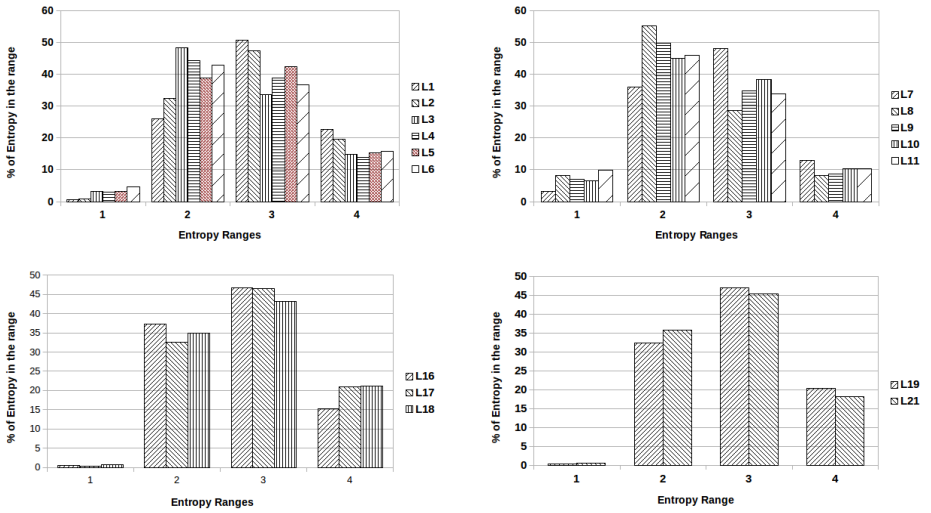

Figure 3: Entropy distribution for considered language pairs

it generates may be different for different languages as well as for different speakers. We perform entropy analysis on the text generated by different synthesizers across different languages. Table 3 presents different language pair where first and third column represent the representative symbol for respective language pair. We investigate distribution of unigram word across languages using Entropy measure.

Entropy is the measure of uncertainty and for each word or term t, it is defined as:

$$H(t) = -\sum_{l=1}^{L} p_l(t) log(p_l(t)) \tag{1}$$

where L is the set of languages and $p_l(t)$ is the probability of the term t in the language l. Fig 3 represents the percentage of entropy distribution for different pair of language combinations (x-axis corresponds to different entropy ranges and y-axis corresponds to percentage of total number of features (terms) that are present in a particular range). We have considered four different ranges; [1,0.1] denoted as 1, (0.1,0.01] denoted as 2, (0.01,0.001] denoted as 3, and (0.001,0.0001] denoted as 4. The entropy of a term (feature) is minimal (H(t)=0), if the term occurs only in one language. From the Fig 3, it is evident that, for all language combinations, 70% of the words lie in lower range (very close to 0). This indeed implies that texts generated by different transcribers are discriminative in nature.

Table 4: Accuracy of different classifiers over the transcription text generated using unbiased synthesizers using Studio Recorded speech samples. The synthesizer S1 denotes Web Speech API synthesizer, S2 denotes IBM synthesizer and S3 denotes Locally built HMM synthesizer

L	RF			DT			SVM			NB1			NB2		
	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3
L1	0.74	0.84	1	0.72	0.84	1	0.81	0.94	1	0.84	0.96	1	0.7	0.83	0.90
L2	0.84	0.87	0.96	0.82	0.82	0.93	0.89	0.94	0.97	0.91	0.96	0.95	0.88	0.9	0.73
L3	0.92	0.89	0.91	0.9	0.87	0.82	0.95	0.96	0.88	0.95	0.98	0.91	0.85	0.9	0.74
L4	0.85	0.89	0.99	0.86	0.89	0.98	0.92	0.97	0.99	0.95	0.99	0.98	0.8	0.93	0.93
L5	0.84	0.93	0.96	0.82	0.91	0.94	0.91	0.97	0.97	0.94	0.98	0.95	0.86	0.86	0.57
L6	0.85	0.93	0.99	0.84	0.9	0.98	0.92	0.97	0.99	0.94	0.98	0.99	0.84	0.88	0.97
L7	0.79	0.85	1	0.84	0.88	0.99	0.9	0.96	1	0.93	0.98	1	0.89	0.92	1
L8	0.91	0.86	0.99	0.87	0.79	0.97	0.94	0.92	1	0.94	0.94	1	0.88	0.82	0.99
L9	0.84	0.88	1	0.82	0.82	1	0.89	0.94	1	0.91	0.96	1	0.82	0.84	0.99
L10	0.8	0.9	1	0.76	0.82	1	0.88	0.96	1	0.92	0.98	1	0.85	0.88	0.99
L11	0.81	0.9	1	0.78	0.82	0.99	0.89	0.96	1	0.93	0.98	1	0.86	0.88	0.97
L12	0.94	0.88	0.94	0.93	0.88	0.9	0.96	0.96	0.94	0.96	0.97	0.98	0.93	0.92	0.91
L13	0.83	0.84	0.99	0.87	0.87	0.99	0.93	0.97	0.99	0.95	0.97	0.99	0.92	0.93	0.96
L14	0.82	0.92	0.92	0.84	0.9	0.89	0.9	0.97	0.93	0.94	0.97	0.9	0.9	0.92	0.83
L15	0.85	0.91	0.99	0.86	0.88	0.99	0.92	0.97	1	0.94	0.97	1	0.92	0.93	0.98
L16	0.87	0.86	0.98	0.85	0.81	0.95	0.94	0.94	0.99	0.94	0.97	0.93	0.86	0.83	0.85
L17	0.8	0.89	0.98	0.76	0.83	0.97	0.88	0.96	0.98	0.92	0.98	0.98	0.85	0.91	0.92
L18	0.88	0.9	1	0.84	0.83	1	0.93	0.96	1	0.96	0.98	0.99	0.92	0.92	0.96
L19	0.94	0.9	0.97	0.93	0.83	0.93	0.96	0.97	0.97	0.98	0.99	0.98	0.95	0.88	0.93
L20	0.81	0.85	1	0.78	0.79	0.99	0.89	0.94	0.99	0.92	0.96	1	0.85	0.85	0.97
L21	0.94	0.88	0.98	0.92	0.81	0.98	0.96	0.95	0.99	0.97	0.97	0.99	0.93	0.86	0.92

From the above text analysis, it is quite evident that, the transcribed text generated by the unbiased or language independent transcribers (LIT) are highly discriminative in nature. It means, the transcribed text may be effectively used to identify underlying language of the speech samples. It also indicates that the transcribers are able to capture intra-language homogeneous characteristics and inter-language discriminating characteristics.

6. Experimental investigation and results

To investigate the performance of various classification methods (ref. Section 3) over the transcription text generated from the unbiased transcribers and compare the performances with their counterparts; (i) using correctly transcribed text, and (ii) audio features, we have setup with various classification frameworks. First, performance of traditional classifiers over the text generated from three individual transcribers and decision fusion based ensembling method are investigated. Second, to understand the effectiveness of the proposed framework over the performance of the traditional classifiers with correctly transcribed text. Further, we also investigate the performance of the classifiers over audio features and compare with that of the proposed framework. To extract audio features, OpenEAR (Eyben et al. 2009) is used. A total of 988 features are extracted for each speech sample using OpenEAR. To observe the impact of length of audio samples in LID, we also considered text generated from speech samples having less than 10 sec duration and 30 sec duration.

6.1. Performance of Base Classifiers

Table 4 shows the performance of traditional classifiers over studio recorded data set considering textual features generated from text obtained from speech samples using above mentioned three transcribers. It is evident from the table that, considering phone level features (with transcribers S3) base classifiers like Random Forest, Decision Tree (ID3), SVM with linear kernel

Table 5: Accuracy for Decision Diffusion Based Ensembling (Figure 2) with S1, S2 and S3 synthesizers

	RF	DT	KNN	SVM	NB1	NB2		RF	DT	KNN	SVM	NB1	NB2
L1	0.99	0.69	1	1	0.97	1	L12	0.89	0.93	0.93	0.96	0.76	0.93
L2	0.94	0.87	0.99	0.95	0.96	0.97	L13	0.99	0.81	0.99	0.99	1	1
L3	0.87	0.88	0.82	0.85	0.78	0.84	L14	0.93	0.91	0.98	0.96	0.87	0.95
L4	1	0.96	1	0.99	0.99	1	L15	0.98	0.99	1	0.99	0.99	1
L5	0.95	0.87	1	0.93	0.83	0.97	L16	0.92	0.82	0.84	0.80	0.76	0.84
L6	0.84	1	1	0.98	0.97	1	L17	0.94	0.94	1	1	1	1
L7	1	1	1	1	1	1	L18	0.99	0.93	1	1	0.98	1
L8	0.93	1	0.94	0.98	0.76	0.96	L19	0.96	0.87	0.96	0.95	0.76	0.96
L9	0.89	1	1	1	1	1	L20	0.94	0.95	0.99	0.95	0.99	0.97
L10	0.99	1	1	1	0.96	1	L21	0.99	0.98	0.97	0.96	0.76	0.97
L11	0.98	0.97	1	0.98	0.96	1							

and Naive Bayes Multinomial can achieve 100% accuracy for language combinations like L9 (Bengali and Gujarati) and L10 (Bengali and Tamil). Further, results in the table also show an interesting observation that text generated using Google (S1) and IBM (S2) synthesizers (which generate word level patterns) provide comparable classification accuracy while the classification over the text generated using phone level synthesizer (S3) provides relatively better accuracy. Transcriber with phone level features is able to achieve an average accuracy of 94% while word level transcribers (S1 and S2) achieve average accuracy of 83% and 86% respectively. Considering the simplicity of the framework (using unbiased transcribers, without having language dependent ASR), an average accuracy of 94% is indeed a promising result.

6.1.1. Performance of DFBE classification framework

Table 5 show the performance of the proposed ensembling framework over various language combinations. For language combination Bengali and Manipuri (L7), we able to achieve an absolute 100% accuracy considering Random Forest, Decision Tree, KNN, SVM Linear, Naive Bayes Multinomial and Naive Bayes Gaussian as second level classifiers. Out of 126 combinations, for 33 combinations, we have achieved 100% identification accuracy. DFBE framework outperforms its counterparts (individual classifiers) in 57.1% of the language pairs and performs equally in 28% of the language pairs. It demonstrates the effectiveness of the decision fusion in language identification. Table 9 presents performance of different classifiers considering all eight languages together.

6.1.2. Performance based on audio length

To investigate the effect of length (duration) of the speech samples used for generating transcription, table 6 compares performance of different classifiers built over transcription text generated from speech samples of different lengths. It also investigates the effect of audio length on classification accuracy. In this table, T1 represents samples having duration less than 10sec whereas T2 represents audio samples having duration between 10sec to 30sec. This study is done on noisy, multi-speaker youtube speech samples. It is observed that text obtained from small duration speech samples outperform text generated from longer duration speech samples for all language pairs. It may be due to the fact that the effect of background noise while generating transcription text is lesser in the case of short duration speech sample.

Table 6: LID Accuracy comparison based on audio length

Lang	RF		DT		SVM		NB1		NB2		RNN	
	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
L1	0.73	0.62	0.72	0.59	0.74	0.63	0.71	0.65	0.62	0.57	0.85	0.71
L2	0.66	0.67	0.64	0.65	0.67	0.68	0.68	0.68	0.63	0.55	0.81	0.65
L3	0.77	0.77	0.76	0.76	0.78	0.77	0.79	0.78	0.69	0.74	0.86	0.76
L4	0.75	0.73	0.74	0.73	0.76	0.75	0.77	0.76	0.57	0.64	0.88	0.73
L5	0.77	0.78	0.75	0.77	0.78	0.79	0.78	0.79	0.54	0.55	0.88	0.77
L6	0.73	0.77	0.72	0.76	0.76	0.79	0.77	0.80	0.57	0.58	0.82	0.76
L7	0.74	0.71	0.76	0.71	0.76	0.73	0.76	0.74	0.64	0.59	0.90	0.75
L8	0.86	0.76	0.85	0.72	0.87	0.77	0.84	0.78	0.79	0.69	0.83	0.87
L9	0.83	0.72	0.83	0.71	0.85	0.74	0.82	0.75	0.67	0.58	0.83	0.81
L10	0.87	0.67	0.86	0.65	0.88	0.83	0.86	0.84	0.64	0.56	0.88	0.85
L11	0.84	0.82	0.83	0.80	0.86	0.82	0.82	0.82	0.67	0.59	0.81	0.79
L12	0.81	0.76	0.79	0.76	0.82	0.77	0.79	0.78	0.78	0.76	0.81	0.79
L13	0.78	0.71	0.78	0.70	0.81	0.75	0.79	0.73	0.71	0.70	0.72	0.7
L14	0.79	0.66	0.79	0.64	0.81	0.74	0.78	0.74	0.62	0.55	0.71	0.68
L15	0.67	0.70	0.67	0.7	0.79	0.73	0.77	0.73	0.65	0.57	0.7	0.69
L16	0.74	0.76	0.72	0.75	0.76	0.77	0.79	0.79	0.68	0.72	0.75	0.71
L17	0.74	0.67	0.72	0.68	0.77	0.8	0.78	0.81	0.72	0.58	0.76	0.69
L18	0.69	0.77	0.68	0.76	0.73	0.78	0.75	0.78	0.63	0.62	0.76	0.71
L19	0.78	0.75	0.75	0.73	0.81	0.83	0.82	0.84	0.78	0.81	0.82	0.8
L20	0.68	0.65	0.68	0.66	0.71	0.67	0.71	0.69	0.64	0.61	0.79	0.71
L21	0.72	0.80	0.71	0.78	0.78	0.81	0.81	0.83	0.75	0.80	0.81	0.75

Table 7: Comparison of classification accuracies built using manually transcribed text and proposed unbiased transcription text for Assamese-Manipuri-Mizo language pairs. In the table, performance of the best performing classifier is reported. The best performing classifier is noted between brackets.

Language Pairs	Assamese	Manipuri	Mizo
Assamese		0.97(NB)	0.92(NB)
Manipuri	0.99(NB)		0.91(NB)
Mizo	0.99(NB)	0.99(NB)	

6.1.3. Performance over manually transcribed text

Manually transcript generation from speech samples is an expensive operation. However, to understand the effectiveness of the proposed framework over correctly transcribed text, we have manually transcribed speech samples for three languages; Assamese, Manipuri and Mizo. Table 7 summarizes the performance of different classifiers (only the classifiers reported in Section 3.1 are considered). The entries in the lower part of the diagonal cells show the best performing classifier for chosen language pairs on manually transcribe text. The entries in the upper diagonal cells show the performance of the same classifier over the proposed transcription text. It is evident from these observations that the proposed transcription framework can provide comparable performance. Considering the simplicity of the framework (using unbiased synthesizers, without language dependent synthesizers), an accuracy of 97% for Assamese-Manipuri pairs using proposed transcribed text as compared to 99% using manually transcribed text is indeed an encouraging result.

Table 8: LID Accuracy comparison considering Audio features and Textual features

Lang	RF		DT		KNN		SVM		NB1		NB2	
	Audio	Text	Audio	Text	Audio	Text	Audio	Text	Audio	Text	Audio	Text
L1	0.99	0.99	0.97	0.99	0.99	0.70	0.98	1	0.52	0.99	0.97	0.993
L2	0.98	0.95	0.97	0.93	0.98	0.92	0.98	0.96	0.51	0.95	0.97	0.73
L3	0.95	0.91	0.93	0.81	0.96	0.8	0.94	0.87	0.71	0.90	0.90	0.74
L4	0.97	0.99	0.96	0.98	0.97	0.98	0.96	0.99	0.6	0.97	0.95	0.93
L5	0.93	0.96	0.89	0.93	0.95	0.89	0.92	0.97	0.62	0.94	0.84	0.57
L6	0.95	0.99	0.92	0.98	0.95	0.96	0.93	0.99	0.69	0.99	0.90	0.96
L7	0.99	1	0.99	0.99	0.99	0.77	0.99	1	0.61	1	0.98	0.99
L8	0.98	0.98	0.98	0.97	0.98	0.93	0.98	1	0.72	0.99	0.98	0.98
L9	0.99	1	0.99	0.99	0.99	0.96	0.99	1	0.62	0.99	0.98	0.98
L10	0.97	1	0.94	0.99	0.97	0.71	0.97	1	0.65	0.99	0.96	0.99
L11	0.94	0.99	0.91	0.98	0.95	0.68	0.93	0.99	0.71	0.99	0.89	0.97
L12	0.99	0.94	0.99	0.89	0.99	0.79	0.99	0.94	0.64	0.97	0.99	0.91
L13	0.95	0.99	0.93	0.98	0.94	0.97	0.93	0.99	0.72	0.99	0.93	0.95
L14	0.90	0.92	0.86	0.88	0.88	0.87	0.84	0.93	0.61	0.89	0.84	0.83
L15	0.96	0.99	0.94	0.98	0.96	0.98	0.95	0.99	0.71	0.99	0.92	0.98
L16	0.99	0.97	0.99	0.95	0.99	0.95	0.99	0.98	0.53	0.93	0.99	0.85
L17	0.98	0.98	0.97	0.96	0.99	0.85	0.98	0.98	0.57	0.97	0.97	0.92
L18	0.99	1	0.98	0.99	0.99	1	0.99	0.99	0.58	0.98	0.99	0.96
L19	0.91	0.96	0.87	0.93	0.90	0.75	0.86	0.97	0.61	0.98	0.83	0.93
L20	0.93	0.99	0.89	0.99	0.91	0.97	0.89	0.99	0.52	0.99	0.91	0.96
L21	0.94	0.98	0.91	0.98	0.92	0.98	0.90	0.98	0.6	0.99	0.91	0.92

Table 9: Performance of classifiers considering eight languages together

	DT	SVM	NB1
Eight Languages	0.735	0.801	0.768

6.1.4. Performance over audio features

We further compare the performance of the proposed framework with that of the audio features using the same classification framework. Table 8 presents the observations using audio features. Over all, over the studio recorded and Youtube data sets, textual feature-based classifiers outperform it's audio counterpart in 76% and 100% cases respectively. We can see that maximum accuracy achieves using audio features is 99% while using textual features we are able to achieve upto 100% accuracy for six (6) language combinations for different classifiers. Average accuracy of 94.7% is achieved considering textual features whereas only an average accuracy of 89.8% is achieved considering speech features.

7. Conclusions and Future Work

In this paper, we have investigated the effectiveness of using language independent speech-to-text transcribers for automatic language identification task over eight Indian languages. From various experimental results, it is evident that an accuracy as high as 99% can be achieved using transcription texts from unbiased (language independent) transcribers. It is also evident that proposed LID framework can achieve comparable accuracy with that of the manually transcribed text and audio features.

At present this paper has considered only one speaker (in case of studio recording) with a particular gender (male). It will be interesting to investigate consistency in the classification performance over multiple speakers of different genders, different age group etc. However, we left this study as future extension.

References

Balleda, J., Murthy, H. A. and Nagarajan, T., 2000, Language identification from short segments of speech, in *Sixth International Conference on Spoken Language Processing*.

Cimarusti, D. and Ives, R., 1982, Development of an automatic identification system of spoken languages: Phase i, in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82.*, vol. 7, pp. 1661--1663, IEEE.

Cummins, F., Gers, F. and Schmidhuber, J., 1999, Comparing prosody across many languages, *Instituto Dalle Molle di Studie sull'Intelligenza Artificiale, Lugano, Switzerland, Tech. Rep., IDSIA-07-99*.

Eyben, F., Wöllmer, M. and Schuller, B., 2009, Openear—introducing the munich open-source emotion and affect recognition toolkit, in *Affective computing and intelligent interaction and workshops, 2009. ACII 2009. 3rd international conference on*, pp. 1--6, IEEE.

Hazen, T. J. and Zue, V. W., 1997, Segment-based automatic language identification, *The Journal of the Acoustical Society of America*, vol. 101, no. 4, pp. 2323--2331.

House, A. S. and Neuburg, E. P., 1977, Toward automatic identification of the language of an utterance. i. preliminary methodological considerations, *The Journal of the Acoustical Society of America*, vol. 62, no. 3, pp. 708--713.

Hughes, B., Baldwin, T., Bird, S., Nicholson, J. and MacKinlay, A., 2006, Reconsidering language identification for written language resources.

Jothilakshmi, S., 2010, Speech analysis for speaker diarization and spoken language identification.

Jothilakshmi, S., Ramalingam, V. and Palanivel, S., 2012, A hierarchical language identification system for indian languages, *Digital Signal Processing*, vol. 22, no. 3, pp. 544--553.

Khan, A., Baharudin, B. and Khan, K., 2010, Semantic based features selection and weighting method for text classification, in *Information Technology (ITSim), 2010 International Symposium in*, vol. 2, pp. 850--855, IEEE.

Li, K.-P., 1994, Automatic language identification using syllabic spectral features, in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, vol. 1, pp. I--297, IEEE.

Mary, L. and Yegnanarayana, B., 2004, Autoassociative neural network models for language identification, in *International conference on intelligent sensing and information processing*, pp. 317--320.

Rish, I. et al., 2001, An empirical study of the naive bayes classifier, in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, pp. 41--46, IBM New York.

Safitri, N. E., Zahra, A. and Adriani, M., 2016, Spoken language identification with phonotactics methods on minangkabau, sundanese, and javanese languages, *Procedia Computer Science*, vol. 81, pp. 182--187.

Saikia, R., Singh, S. R. and Sarmah, P., 2017, Effect of language independent transcribers on spoken language identification for different indian languages, in *Asian Language Processing (IALP), 2017 International Conference on*, pp. 214--217, IEEE.

Skowron, A., Wang, H., Wojna, A. and Bazan, J., 2006, Multimodal classification: case studies, in *Transactions on Rough Sets V*, pp. 224--239, Springer.

Sutskever, I., 2013, *Training recurrent neural networks*, University of Toronto Toronto, Ontario, Canada.

Tam, V., Santoso, A. and Setiono, R., 2002, A comparative study of centroid-based, neighborhood-based and statistical approaches for effective document categorization, in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 4, pp. 235--238, IEEE.

Thymé-Gobbel, A. E. and Hutchins, S. E., 1996, On using prosodic cues in automatic lan-

guage identification, in *Fourth International Conference on Spoken Language Processing*.
Vapnik, V., 2013, *The nature of statistical learning theory*, Springer science & business media.
Yan, Y. and Barnard, E., 1995, An approach to automatic language identification based on language-dependent phone recognition, in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 5, pp. 3511--3514, IEEE.