ISIS and NISIS: New Bilingual Dual-Channel Speech Corpora for Robust Speaker Recognition

Amita Pal[1], Smarajit Bose[1], Mandar Mitra[2] and Sandipan Roy[3]
[1]BIRU, Indian Statistical Institute, Kolkata, India
[2]CVPRU, Indian Statistical Institute, Kolkata, India
[3]Department of Statistics, University of Michigan, Ann Arbor, MI, USA

Abstract – *It is standard practice to use benchmark datasets for comparing meaningfully the performance of a number of competing speaker identification systems. Generally, such datasets consist of speech recordings from different speakers made at a single point of time, typically in the same language. That is, the training and test sets both consist of speech recorded at the same point of time in the same language over the same recording channel. This is generally not the case in real-life applications.*

In this paper, we introduce a new database consisting of speech recordings of 105 speakers, made over four sessions, in two languages and simultaneously over two channels. This database provides scope for experimentation regarding loss in efficiency due to possible mismatch in language, channel and recording session. Results of experiments with MFCC-based GMM speaker models are presented to highlight the need of such benchmark datasets for identifying robust speaker identification systems.

Keywords: Robust speaker recognition, mel frequency cepstral coefficients, Gaussian mixture models, classification accuracy

1 Introduction

In speaker recognition research, as in other areas of applied research, large databases are indispensible for uniform evaluation of emerging methodologies relative to competing or existing ones.

Most popular databases used for speaker recognition contain a number of utterances of several speakers recorded in a single session in a specific language over a single channel. For instance, TIMIT [1,2] is a corpus of read English text with 630 speakers, which is very widely used. NTIMIT [3,4] and CTIMIT [5] are noisier versions of TIMIT, created after the voice samples in TIMIT were passed through telephone and cellular phone channels respectively.

Speaker recognition has significant commercial and forensic applications, where only those approaches which ensure robustness with respect to variability in language/dialect, (recording) session, recording channel, and so on, find acceptability. This concern stems from the fact that in such real-life applications, test samples are usually compared with training samples recorded at an earlier point of time, that is, in an earlier session. Since the voice characteristics of a speaker do vary over time, the inter-session variability between the training and test samples degrades the performance of a speaker recognition method. Moreover, the test sample and the training samples are generally not in the same language or dialect, and may also be recorded over different channels. These factors also affect the performance of speaker recognition systems adversely, just as they affect the characteristics of a voice sample.

Hence it is interesting to explore how the accuracy of existing methods gets affected by this additional variability induced by differences in languages, recording channels and sessions between training and test data. For this purpose, a database with voice samples of speakers recorded at different time points, in different languages, simultaneously through a number of channels, is required.

ISIS (an acronym for Indian Statistical Institute Speech) and NISIS (Noisy ISIS) are precisely such speech corpora, which contain simultaneously recorded microphone and telephone speech respectively, over multiple sessions, spontaneous as well as read, in two languages (Bangla and English), recorded in a typical office environment with moderate background noise. They were created in the Indian Statistical Institute, Kolkata, as a part of a project funded by the Department of Information Technology, Ministry of Communications and Information Technology, Government of India, during 2004-07. Details of the methodology and the database are given in subsequent sections. The speakers generally had Bangla or another Indian language as their mother tongue, and so were non-native English speakers.

The paper is organized as follows. Section 2 describes the ISIS and NISIS databases. To highlight the features of the database, GMM-MFCC classification [6] was used for non-contemporary speaker identification with training and test sets from same as well as different languages. The results of experiments with GMM-MFCC classifiers on these datasets

are given in Section 3. It clearly shows that the accuracy deteriorates to some extent as time passes by. The accuracy is also affected if the languages in the training and test samples differ. In Section 4, we illustrate how some modified GMM-MFCC classifiers can improve the performance significantly when there is a difference in time or language. Finally, Section 5 provides concluding remarks and future directions.

2 Description of the Database

Particulars of the Bangla and English databases in ISIS are given below:

- Number of speakers : 105 (53 male + 52 female)
- Recording environment: moderately quiet computer room
- Sessions per speaker: 4
- Interval between sessions: 1 week to about 2 months
- Types of utterances:
  - 10 isolated words (randomly drawn from a specific text corpus described below, and generally different for all speakers and sessions)
  - answers to 8 questions (these answers included dates, phone numbers, alphabetic sequences, and a few words spoken spontaneously)
  - 12 sentences (first two sentences common to all speakers, the remaining randomly drawn from the text corpus, duration ranging from 3-10 seconds)

A total of 122 volunteers attended at least one recording session. However, 17 of them attended fewer than 4 sessions. The recordings of these sessions are also available separately.

For each volunteer, information about his/her gender, age, and educational qualifications are available. In addition, some information about where the person had grown up is also stored.

2.1 Methodology Used

Each session took place in a moderately quiet computer room. In each session, a volunteer recorded utterances simultaneously via two channels:

- a lapel microphone (Ahuja UTP-30), connected to a Kay Computerized Speech Lab (CSL 4500)
- an intercom telephone connection via a standard telephone headset (Panasonic KX-TCA86) attached to a cordless phone (Panasonic KX-TG2448BX), connected to the intercom system at ISI. Calls were received and recorded using an Intel dialogic card (model D/41 JCT-LS).

Since Kay CSL 4500 is Windows-based, it was installed on a desktop computer with Windows OS. On the other hand, since our dialogic card is LINUX-based, it was installed on another desktop computer with LINUX OS. Recordings were done simultaneously with these systems, using C programs and macros. A front-end was designed to make each recording session interactive, by displaying the text to be read, with manual control of the beginning and end of each recording, according to the convenience and readiness of the subject. Figure 1 provides a schematic representation of the system setup used for each recording session.

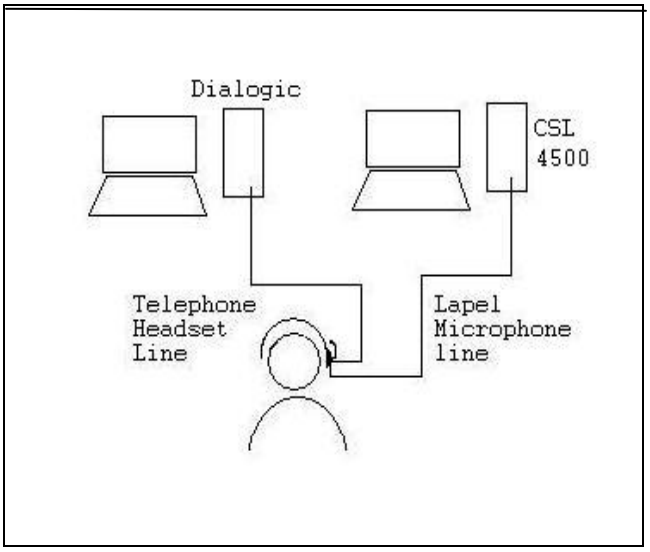

Figure 1: Schematic view of the recording setup

2.2 Text Corpus

The text used in the recording of the English language database is entirely taken from the TIMIT text corpus. On the other hand, the text used for the Bangla language database is taken from the Ananda Bazaar Publication corpus. This corpus consists of newspaper and magazine articles.

A total of 500 words and 300 sentences were chosen to build each corpus. Two sentences were initially chosen randomly from either corpus, and were common for all sessions of all speakers. For each recording session of a speaker, 10 distinct words and 10 distinct sentences were randomly selected afresh from the respective corpus. From the last 10 sentences that were not common to all speakers, 8 sentences that were best recorded in terms of high signal-to-noise ratio, were retained in the database. The rejected recordings are also available separately.

2.3 Calibration

For every subject, at each session, a calibration recording (almost 4 sec.) was made over both channels (telephone and microphone), to ensure that the signal-to-noise ratio (SNR) was above a threshold of 30 for telephone, and 20 for microphone. This automatic monitoring was done with the help of software specifically written for this purpose, which was run before each recording.

3 Speaker Identification Results

To get an idea about the classification accuracy expected with this database, the Gaussian Mixture Model (GMM) based on Mel Frequency Cepstral Coefficients (MFCCs) as features, proposed by Reynolds [6], was implemented on it. Each utterance is split into overlapping segments or frames of length 256 *ms*, the frame advance being 64 *ms*. From each frame, 38 MFCCs were computed, and a GMM with 32 components based on this 38-dimensional feature vector was used for each speaker.

Only utterances from the "sentences" folder were used and, of the 10 recordings for each speaker at each session, the first 6 were used for training, while the remaining 4 were used as test data.

Table I: Classification Accuracy on ISIS with Different Combinations of Training and Test Sets

Train Set ↓ \ Test Set→		ENGLISH				BANGLA			
		Session 1	Session 2	Session 3	Session 4	Session 1	Session 2	Session 3	Session 4
ENGLISH	Session 1	**94**	79	80	78	86	74	79	75
	Session 2	71	**95**	81	80	62	88	70	78
	Session 3	71	83	**96**	83	57	69	79	75
	Session 4	73	79	82	**95**	53	62	65	85
BANGLA	Session 1	88	66	62	58	**96**	77	72	72
	Session 2	62	87	66	61	71	**95**	73	79
	Session 3	66	73	89	70	69	79	**90**	84
	Session 4	64	76	75	84	71	78	79	**96**

Tables I and II give the classification results for ISIS and NISIS respectively, for different combinations of test and training sets, varying with respect to language and session. As expected, the best results are obtained when both training and test data are in the same language and are contemporary, that is, recorded in the same session, and there is a general deterioration when

Table II: Classification Accuracy on NISIS with Different combinations of Training and Test Sets

Train Set ↓ \ Test Set→		ENGLISH				BANGLA			
		Session 1	Session 2	Session 3	Session 4	Session 1	Session 2	Session 3	Session 4
ENGLISH	Session 1	**68**	40	35	33	33	29	29	25
	Session 2	41	**71**	37	31	34	37	28	27
	Session 3	45	40	**69**	34	34	32	36	28
	Session 4	29	36	38	**64**	27	28	32	34
BANGLA	Session 1	51	41	36	29	**71**	41	42	39
	Session 2	43	45	37	36	47	**70**	45	38
	Session 3	43	42	41	25	36	39	**69**	42
	Session 4	37	39	42	41	45	44	48	**70**

- the training and test data correspond to different languages;
- the training and test data are recorded at different points of time, the degradation in performance becoming more pronounced in general as the degree of non-contemporaneity increases;
- training and test data differ both in language and in session, the drop in accuracy being generally more pronounced in such cases.

As expected, overall the results are poorer in the case of NISIS compared to that of ISIS.

4 Improved Accuracy with MFCC-GMM algorithms

Bose *et al.* [7] showed how to use the principal component transformation to improve classification accuracy in noisy environments. For every speaker model, MFCCs are transformed using the principal component transformation before computing the GMM models. A test utterance is also transformed using the principal component transformation corresponding to the speaker model with which it is matched. Table III presents the results of one such experiment where English Session 1 samples were used as training samples. In this case the last 6 sentences for each speaker were used for training, and the rest were used for testing. It is quite evident that the loss of efficiency is recovered to a large extent in case of language or session mismatch. This is true for both ISIS and NISIS.

Table III: Improvement in accuracy by applying Principal Component Transformation (PCT) (with English Session 1 recordings as training data)

Test Session		% accuracy before applying PCT	% accuracy after applying PCT
ENGLISH	Session 1	94	97.5
	Session 2	79	84.5
	Session 3	80	85.5
	Session 4	77.5	88
BANGLA	Session 1	86	96.5
	Session 2	74	83.5
	Session 3	79	86.5
	Session 4	74.5	83

In Table IV, we present some preliminary results of ensemble classification method using English session 1 samples again as training data. A number of classifiers were built using different parameter combinations in the GMM-MFCC models, and their decisions were combined by pooling their likelihood matching scores. It is quite clear that this aggregate GMM-MFCC classifier can make further improvement in classification accuracy when the performance is degraded due to time or language mismatch.

Table IV: Summary of Results with English Session 1 as Training Data

		Without Applying PCA		**After Applying PCA**	
	Test Session	Language		Language	
		English	Bangla	English	Bangla
Individual	1	98%	95%	94%	85%
	2	82%	80%	73%	68%
After combination	1	99%	96%	95%	88%
	2	85%	87%	76%	72%

5 Conclusion

In this paper, we have introduced ISIS and NISIS, two new bilingual dual-channel speech corpora which has the unique feature of having speech samples of the same speakers recorded simultaneously in two channels and in two languages at different time points. These databases provide an excellent opportunity to study the degradation in classification accuracy due to channel, time-point or language mismatch. Some preliminary results using GMM-MFCC models and its robust variants, we show how the classification accuracy can be recovered to a large extent. These databases can be effectively used in future development of robust speaker identification algorithms.

References

[1] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, and V. Zue. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Linguistic Data Consortium, Philadelphia.

[2] V. Zue, S. Seneff, and J. Glass. (1990). "Speech database development at MIT: TIMIT and beyond," *Speech Commun*. vol. 9,pp. 351-356.

[3] W.M. Fisher, G.R. Doddington, K.M. Goudie-Marshall, C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz. (1993). *NTIMIT*. Linguistic Data Consortium, Philadelphia.

[4] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz. (1990). "NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database," *International Conference on Acoustics, Speech, and Signal Processing, 1990* (*ICASSP-90*).

[5] K.L. Brown, and E.B. George. (1995). "CTIMIT: a speech corpus for the cellular environment with applications to automatic speech recognition," *International Conference on Acoustics, Speech, and Signal Processing, 1995* (*ICASSP-95*).

[6] D.A. Reynolds, and R.C. Rose. (1995). "Robust Text-Independent Speaker Identification Using Gaussian Mixture Models," *IEEE Transactions on Speech and Audio Processing, vol.* 3, pp. 72-83.

[7] S. Bose, A. Pal, D. Sengupta, and G.K. Basak. (2012). "Improved Speaker Identification with Gaussian Mixture Models (GMMs)," in *Proceedings of the 3rd ICSIIT 2012, International Conference on Soft Computing, Intelligent System and Information Technology*, Bali, Indonesia, 24-25 May 2012 (to appear).