A Multilingual Speech Database for Speaker Recognition

Utpal Bhattacharjee
Department of Computer Science and Engineering,
Rajiv Gandhi University, Rono Hills, Doimukh,
Arunachal Pradesh, India, Pin-791112
Email: utpalbhattacharjee@rediffmail.com,

Kshirod Sarmah
Department of Computer Science and Engineering,
Rajiv Gandhi University, Rono Hills, Doimukh,
Arunachal Pradesh, India, Pin-791112
Email: kshirodsarmah@gmail.com

Abstract–**This paper report the experiments carried out on the recently collected speaker recognition database to study the impact of language variability on speaker verification system. The speech database consists of speech data recorded from 100 speakers with Arunachali languages of North-East India as mother tongue. The speech data is collected in three different languages English, Hindi and a local language of Arunachal Pradesh. The collected database is evaluated with Gaussian mixture model based speaker verification system. The impact of the mismatch in training and testing language has been evaluated. The initial study explores the impact of language mismatched in the training and testing on the performance of the speaker verification system.**

I. INTRODUCTION

The speaker verification system aims to verify whether an input speech corresponds to the claimed identity or not. A security system based on this ability has great potential in several application domains. Speaker verification systems are typically distinguished into two categories – text-dependent and text-independent [1]. In text-dependent system, a predetermined group of words or sentences is used to enroll the speaker to the system and those words or sentences are used to verify the speaker. Text-dependent system use an explicit verification protocol, usually combined with pass phrases or Personal Identification Number (PIN) as an additional level of security. In text-independent system, no constraints are placed on what can be said by the speaker. It is an implicit verification process where the verification is done while the user is performing some other tasks like talking with the customer care executive or registering a complain.

The state-of-art speaker verification system use either adaptive Gaussian mixture model (GMM) [2] with universal background model (UBM) or support vector machine (SVM) over GMM super-vector [3]. Mel-frequency Cepstral coefficients are most commonly used feature vector for speaker verification system. Supra-segmental features like – prosody, speaking style are also combined with the cepstral feature to improve the performance[4].

Till date, most of the speaker verification system operates only in a single-language environment. For a highly multilingual country like India, the effect of multiple languages on state-of-art speaker verification system needs to be investigated. Most of the publicly available databases for speaker verification research are developed in western context, which is not suitable for evaluating the performance of the system in Indian context. Further, the linguistic scenario of North-East India is different from the rest of India. This is the region where two major linguistic families - Indo-European and Tibeto-Burman meet together and speak each others' language fluently.

To evaluate the speaker verification system in multi-lingual environment, a multi-lingual speaker recognition database has been developed and initial experiments were carried out to evaluate the impact of language variability on the performance of the baseline speaker verification system.

The rest of the paper is organized as follows: Section–II describes the details of the speaker recognition database. Section–III details the speaker verification system. The experimental setup, data used in the experiments and result obtained are described in Section IV. The paper is concluded in Section–V.

II. SPEAKER RECOGNITION DATABASE

In this section we describe the recently collected Arunachali Language Speech Database (ALS-DB). Arunachal Pradesh of North East India is one of the linguistically richest and most diverse regions in all of Asia, being home to at least thirty and possibly as many as fifty distinct languages in addition to innumerable dialects and subdialects thereof [5]. The vast majority of languages indigenous to modern-day Arunachal Pradesh belong to the Tibeto-Burman language family. The majority of these in turn belong to a single branch of Tibeto-Burman, namely Tani. Almost all Tani languages are indigenous to central Arunachal Pradesh while a handful of Tani languages are also spoken in Tibet. Tani languages are noticeably characterized by an overall relative uniformity, suggesting relatively recent origin and dispersal within their present-day area of concentration. Most Tani languages are mutually intelligible with at least one other Tani language, meaning that the area constitutes a dialect chain. In

978-1-4673-1318-6/12/$31.00 ©2012 IEEE

addition to these non-Indo-European languages, the Indo-European languages Assamese, Bengali, English, Nepali and especially Hindi are making strong inroads into Arunachal Pradesh primarily as a result of the primary education system in which classes are generally taught by immigrant teachers from Hindi-speaking parts of northern India. Because of the linguistic diversity of the region, English is the only official language recognized in the state.

To study the impact of language variability on speaker recognition task, ALS-DB is collected in multilingual environment. Each speaker is recorded for three different languages – English, Hindi and a local language, which belongs to any one of the four major Arunachali languages - Adi, Nyishi, Galo and Apatani. Each recording is of 4-5 minutes duration. Speech data were recorded in parallel across four recording devices, which are listed in table -1.

TABLE 1: DEVICE TYPE AND RECORDING SPECIFICATIONS

Device Sl. No	Device Type	Sampling Rate	File Format
1	Table mounted microphone	16 kHz	wav
2	Headset microphone	16 kHz	wav
3	Laptop microphone	16 kHz	wav
4	Portable Voice Recorder	44.1 kHz	mp3

The speakers are recorded for reading style of conversation. The speech data collection was done in laboratory environment with air conditioner, server and other equipments switched on. The speech data was contributed by 52 male and 48 female informants chosen from the age group 20-50 years. During recording, the subject was asked to read a story from the school book of duration 4-5 minutes in each language for twice and the second reading was considered for recording. Each informant participates in four recording sessions and there is a gap of at least one week between two sessions.

III. EXPERIMENTAL SETUP

A speaker verification system was developed using Gaussian Mixture Model with Universal Background model (GMM-UBM) based modeling approach. A 38-dimensional feature vector was used, made up of 19 mel-frequency cepstral coefficient (MFCC) and their first order derivatives. The first order derivatives were approximated over three samples. The coefficients were extracted from a speech sampled at 8 KHz with 16 bits/sample resolution. A pre-emphasis filter $H(z)=1-0.96z^{-1}$ has been applied before framing. The pre-emphasized speech signal is segmented into frame of 20 ms with frame frequency 100 Hz. Each frame is multiplied by a Hamming window. From the windowed frame, FFT has been computed and the magnitude spectrum is filtered with a bank of 20 triangular filters spaced on Mel-scale and constrained into a frequency band of 300-3400 Hz. The log-compressed filter outputs are converted to cepstral coefficients by DCT. The 0^{th} cepstral coefficient is not used in the cepstral feature vector since it corresponds to the energy of the whole frame [6], and only 19 MFCC coefficients have been used. To capture the time varying nature of the speech signal, the first order derivative of the Cepstral coefficients are also calculated. Combining the MFCC coefficients with its first order derivative, we get a 38-dimensional feature vector. Cepstral mean subtraction has been applied on all features to reduce the effect of channel mismatch.

The Gaussian mixture model with 1024 Gaussian components has been used for both the UBM and speaker model. The UBM was created by training the speaker model with speaker's data with Expectation Maximization (EM) algorithm and finding the average of all these models [7]. The speaker models were created by adapting only the mean parameters of the UBM using maximum a posteriori (MAP) approach with the speaker specific data.

The detection error trade-off (DTE) curve has been plotted using log likelihood ratio between the claimed model and the UBM and the equal error rate (EER) obtained from the DTE curve has been used as a measure for the performance of the speaker verification system.

IV. EXPERIMENTS

All the experiments reported in this paper are carried out using the database described in section – II. The speech material is first downsampled at 8 KHz. An energy based silence detector is used to identify and discard the silence frames prior to feature extraction. Only data from the headset microphone has been considered in the present study. All the four available sessions were considered for the experiments. Each speaker model was trained using one complete session. The test sequences were extracted from the next three sessions. In the next iteration, the sessions changed their role in a circular fashion. Finally, the mean accuracy among all iterations was calculated as final result. The training set consists of speech data of length 120 seconds per speaker. The test set consists of speech data of length 15 seconds, 30 seconds and 45 seconds. The test set contains more than 3500 test segments of varying length and each test segment will be evaluated against 11 hypothesized speakers of the same sex as segment speaker [8].

A. *Experiment1*

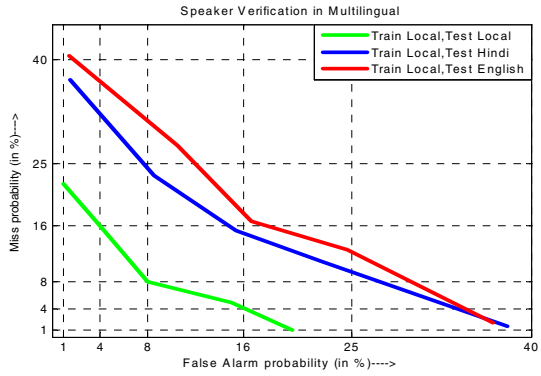
In the first experiment single language has been considered for training the system and each language has been considered separately for testing the system. Training sample of length 120 seconds from a single session has been considered for training the system and the other three sessions have been considered for testing the system. Testing sample of length 15 seconds, 30 seconds and 45 seconds have been extracted from the speech sample of length 120 seconds. The result of the experiments has been summarized in table–2 and table-3. Figure–1 shows the DET curves obtained for the three languages in the speech database.

TABLE-2: EER FOR SPEAKER VERIFICATION SYSTEM FOR TRAINING WITH ONE LANGUAGE AND TESTING WITH EACH LANGUAGE

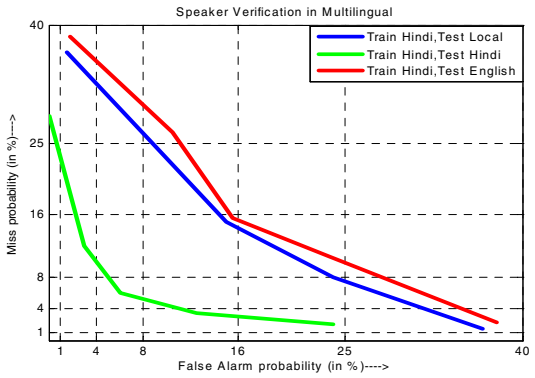
Training Language	Testing Language	FAR %	FRR %	EER
Local	Local	8.0	8.0	8.0
	Hindi	15.3	15.3	15.3
	English	16.6	16.6	16.6
Hindi	Local	15.0	15.0	15.0
	Hindi	6.0	6.0	6.0
	English	15.5	15.5	15.5
English	Local	16.3	16.3	16.3
	Hindi	15.4	15.4	15.4
	English	8.5	8.5	8.5

TABLE 3: ACCURACY OF THE SYSTEM FOR TRAINING IN ONE LANGUE AND TESTING WITH EACH LANGUAGE

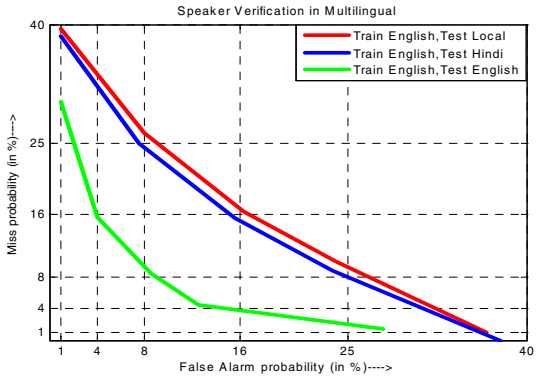
Training Language	Testing Language		
	Local	Hindi	English
Local	92.0%	84.7%	83.4 %
Hindi	85.0%	94.0%	84.5%
English	83.7%	91.5%	91.5%



(a)



(b)



(c)

Figure1. DET curve for the speaker verification system for training with (a) Local (b) Hindi and (c) English and testing with each language.

From the above experiments it has been observed that there is a clear degradation in performance with the change in training and testing language. When the speaker verification system is trained and tested with local language 92% recognition accuracy has been achieved. However, the recognition accuracy drops to 84.7% and 83.4% respectively when the same system is tested with Hindi and English. The same scenario prevails when the system is trained with Hindi and English. When the system is trained and tested with Hindi, the recognition accuracy is 94% whereas in English-English case it is 91.5%. In both the cases there is nearly 8% degradation in performance in training and testing languages mismatched condition.

B. *Experiment2*

In this experiment, combination of two languages has been considered for training the system. The performance of the system has been evaluated using all the three languages separately. Each speaker model has been trained using training data from any two languages of 120 seconds

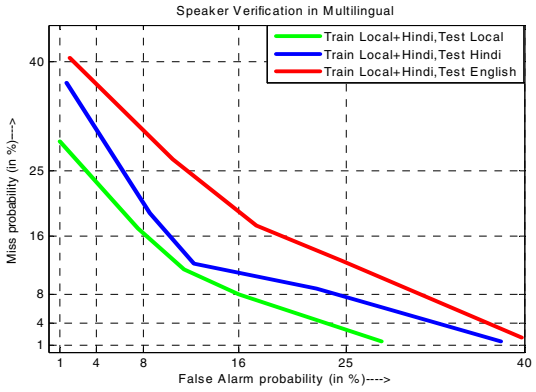
durations each, collected from the same session. All the other sessions have been considered for evaluation of the system. Table-4 and table-5 summarized the results of the experiments and Figure-2 show the DET curve of the experiment.

TABLE 4: EER FOR SPEAKER VERIFICATION SYSTEM FOR TRAINING WITH THE COMBINATION OF TWO LANGUAGES AND TESTING WITH EACH LANGUAGE

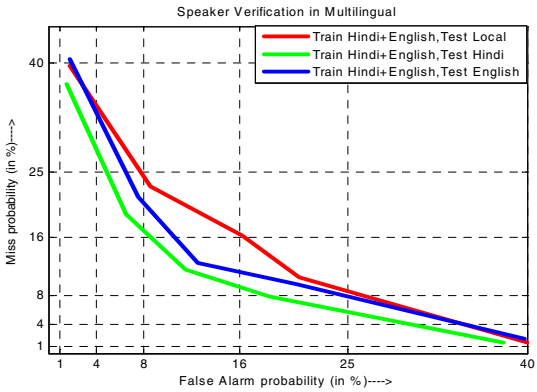
Training Language	Testing Language	FAR %	FRR %	EER
Local + Hindi	Local	11.4	11.4	11.4
	Hindi	12.2	12.2	12.2
	English	17.5	17.5	17.5
Hindi + English	Local	16.2	16.2	16.2
	Hindi	11.5	11.5	11.5
	English	12.5	12.5	12.5
English + Local	Local	11.8	11.8	11.8
	Hindi	15.8	15.8	15.8
	English	10.5	10.5	10.5

TABLE 5: ACCURACY OF THE SYSTEM FOR TRAINING WITH THE COMBINATION OF TWO LANGUAGES AND TESTING WITH EACH LANGUAGE

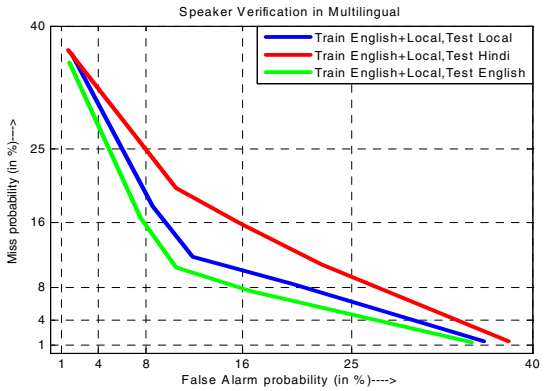
Training Language	Testing Languages		
	Local	Hindi	English
Local + Hindi	88.6%	87.8%	82.5 %
Hindi + English	83.8%	88.5%	87.5%
English + Local	88.2%	84.2%	89.5%



(a)



(b)



(c)

Figure 2: DET curve for the speaker verification system of training with the combination of (a) Local + Hindi (b) Hindi + English and (c) English + Local and testing with each language.

It has been observed that when the system is trained with multiple languages, the system give almost same performance for both the languages used for training. When the system is trained with local + Hindi language, it gives 88.6% and 87.7% recognition accuracy for local and Hindi language test respectively. However, this recognition rate is much lower than the Local-Local and Hindi-Hindi case in previous experiment which is a recognition accuracy of 92% and 94% respectively. The same observations have been made for Hindi + English and English + Local training cases.

C. Experiment3

In this experiment, all the three languages have been considered for training the system. The performance of the system has been evaluated using all the three languages independently. Each speaker model has been trained using training data of duration 120 seconds from each language, collected from the same recording session. The remaining three sessions have been considered for the evaluation of the system. Figure-3 show the DET curve of the experiment and table-6 and table-7 summarized the result.

TABLE 6: EER FOR SPEAKER VERIFICATION SYSTEM FOR TRAINING WITH THE COMBINATION OF ALL THE THREE LANGUAGES AND TESTING WITH EACH THE LANGUAGES SEPARATELY

Training Language	Testing Language	FAR %	FRR %	EER
Local + Hindi + English	Local	18.4	18.4	18.4
	Hindi	17.2	17.2	17.2
	English	17.8	17.8	17.8

TABLE 7: ACCURACY OF THE SYSTEM FOR TRAINING WITH THE COMBINATION OF THREE LANGUAGES AND TESTING WITH EACH LANGUAGE

Training Language	Testing Language		
	Local	Hindi	English
Local + Hindi + English	81.6%	82.8%	82.2 %

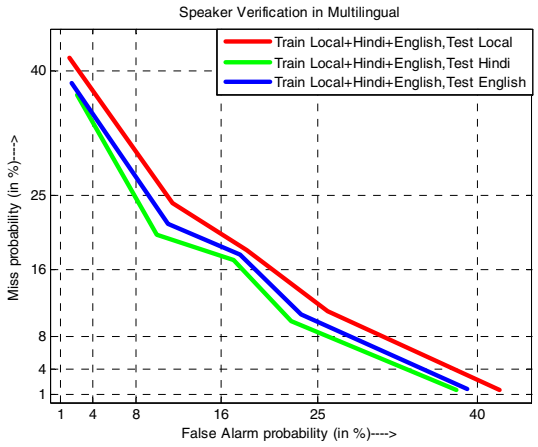


Figure 3: DET curve of the speaker verification system for training with the combination of all the three languages and testing with each language data.

It has been observed that when the system is trained using all the three languages, almost same performance has been observed for each language. However, the observed performance is much degraded when compared to matching conditions of single language experiment and also with matching conditions of two language cases.

V. CONCLUSION

In the present study experiments have been carried out on a recently collected speech database to study the impact of language variability on speaker verification system. Though speech has been considered as biometric, the performance of the speaker verification system found to be highly dependent on the training and the testing languages. This dependency is not unexpected due to the variation of phonetic contents, phonetic constraints and speaking rhythms etc. across different languages. In the present study, we have consider three languages, out of which English and Hindi belongs to Indo-European group and the local language belongs to Tibeto-Burman group. It has been observed that the effects of language variability within the group and outside the group are almost similar. Further, it has been observed that if the system is trained with more than one language, the relative performance of the system degrades in matching conditions compared to that in single language cases. It may be due to the presence of nearly similar phonetic contents in multiple training languages which update the Gaussian components corresponding to those phonemes. Further in-depth study in this direction will reveal many issues which may be significant for language robustness of speaker verification system.

ACKNOWLEDGEMENT

This work has been supported by the ongoing project grant No. 12(12)/2009-ESD sponsored by the Department of Information Technology, Government of India.

REFERENCES

[1] Rosenberg, J. Delong, C. Lee, B. Juang, and F. Soong, "The use of cohort normalized scores for speaker recognition", *In Proc. ICSLP*, pp. 599–602, 1992.

[2] A. Reynolds, "Robust text-independent speaker identification using Gaussian mixture speaker models", *Speech Communications*, vol. 17, pp. 91-108, 1995.

[3] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing*, vol. 10(1–3), pp. 19-41, 2000.

[4] Haris B.C., Pradhan G., Misra A, Shukla S., Sinha R and Prasanna S.R.M., Multi-variability Speech Database for Robust Speaker Recognition, *In Proc. NCC*, pp. 1-5, 2011.

[5] Arunachal Pradesh, http://en.wikipedia.org/wiki/Arunachal_Pradesh

[6] Xiaojia Z., Yang S. and DeLiang W., "Robust speaker identification using a CASA front-end", *In Proc. ICASSP*, 2011 IEEE International Conference on, pp.5468-5471, 2011.

[7] Kleynhans N.T. and Barnard E., Language dependence in multilingual speaker verification, *in Proc. of the 16th Annual Symposium of the Pattern Recognition Association of South Africa*, Langebaan, South Africa, pp. 117-122, 2005.

[8] NIST 2003 Evaluation plan, http://www.itl.nist.gov/iad/mig/tests/sre/2003/2003-spkrec-evalplan-v2.2.