Multi-Variability Speech Database for Robust Speaker Recognition

Haris B C, G Pradhan, A Misra, S Shukla, R Sinha and S R M Prasanna
Department of Electronics and Communication Engineering
Indian Institute of Technology Guwahati, Guwahati -781039, India
Email: {haris, gayadhar, abhinavmisra, sumit.shukla, rsinha, prasanna}@iitg.ernet.in

***Abstract*—In this paper, we present our initial study with the recently collected speech database for developing robust speaker recognition systems in Indian context. The database contains the speech data collected across different sensors, languages, speaking styles, and environments, from 200 speakers. The speech data is collected across five different sensors in parallel, in English and multiple Indian languages, in reading and conversational speaking styles, and in office and uncontrolled environments such as laboratories, hostel rooms and corridors etc. The collected database is evaluated using adapted Gaussian mixture model based speaker verification system following the NIST 2003 speaker recognition evaluation protocol and gives comparable performance to those obtained using NIST data sets. Our initial study exploring the impact of mismatch in training and test conditions with collected data finds that the mismatch in sensor, speaking style, and environment result in significant degradation in performance compared to the matched case whereas for language mismatch case the degradation is found to be relatively smaller.**

I. Introduction

Speaker verification (SV) is a technology which is used to authenticate persons from their voice samples. The state of the art speaker verification systems use either adapted Gaussian mixture models (GMM) with universal background models (UBM) [1] or support vector machines (SVM) over GMM super-vectors [2] for modeling the speakers. Mel-frequency cepstral coefficients (MFCC) are the most commonly used features and are also combined with supra-segmental informations such as prosody and speaking style for improved performance [3]. The main applications of the SV technology are in person authentication and in forensic science. With the growth in the wireless telecommunication, many of these applications are now accessed through mobile phones. In case of mobile phone based access of the SV system it gets exposed to large variation in the handset devices and environmental conditions in addition to channel variability. In a country like India with multilingual community, the system is also expected to work with different languages and accents.

The speaker verification research in recent times is concentrated toward addressing the mismatch between training and testing conditions. Several methods have been developed to address this problem include various normalization techniques [4], feature mapping [5], speaker model synthesis [6], factor analysis [7] and nuisance attribute projection (NAP) [8]. Many of these techniques require the availability of parallel condition data. Amongst the publicly available speaker recognition databases many do not contain parallel data for different conditions. In addition, most of the speaker recognition databases available are in American English and do not cover Indian languages and environmental conditions. To overcome this constraint, we describe the creation of multi-device, multi-lingual and multi-environment speech database for speaker recognition tasks and is referred to as the *IITG Multi-Variability* (IITG-MV) speaker recognition database. We also report our initial study to access the impact of various sensor, language style and environmental mismatch conditions on the baseline speaker verification system developed using this database.

The rest of the paper is organized as follows. Section II describes the details of the MV speaker recognition database. The experimental setup and its performance for the NIST and MV databases are presented in Section III. Section IV describes various experiments conducted on the collected database involving different mismatch conditions. Finally the paper is concluded in Section V.

II. Multi-Variability Speaker Recognition Database

In this section, we describe in detail the recently collected IITG-MV speaker recognition database. To study the impact of different variabilities on the speaker recognition task, the speech data is collected in multi-sensor, multi-lingual, multi-style and multi-environment conditions. The database contains two sets having two recording sessions. Each set contains 100 speakers, with 50 speakers common across both sets. The first set is collected in office-environment involving multiple sensors, multiple languages, and different speaking styles (conversational and read speech) and it is referred to as the IITG-MV Phase-I. The second set differs from the first one in data collected in uncontrolled environments such as laboratories, hostel rooms and corridors etc., while keeping the other variabilities unchanged and is referred to as the MV Phase-II.

A. IITG-MV Phase-I Data set

In this phase the following three types of variabilities were considered while collecting the speech data from different speakers:

- Multi-sensors: Speech data were recorded over five different sensors in parallel.

978-1-61284-091-8/11/$26.00 ©2011 IEEE

Table I: Technical details of the sensors/devices used for collecting the speech data

Device/sensor	Make/model	Sampling Rate	Recording format
Headset mic	Frontech JIL 1903	16 kHz	wav
Tablet PC	HP Elite Book 2730p	16 kHz	wav
Mobile phone-1	Nokia 5130 XpressMusic	8 kHz	amr
Mobile phone-2	Sony Ericsson W350	8 kHz	amr
DVR	Sony ICD-UX70	44.1 kHz	mp3

- Multi-lingua: Every speaker spoke in two different languages: English and his/her favorite Indian language.
- Multi-styles: Every speaker spoke in reading and conversational styles.

In Phase-I, the speech data collection was done in a small office room with electric fan and air conditioner switched on. The data was collected in parallel with a headset microphone connected to a Tablet PC, the built-in microphone of another Tablet PC, two mobile phones of different make with voice recording facility and one digital voice recorder (DVR). The motivation behind choosing these devices for data recording was to get a good sample of the sensors commonly used in portable devices. The technical details of these sensors/devices are summarized in the Table I.

The speech data was contributed by 81 male and 19 female subjects chosen from the student, staff and faculty community at Indian Institute of Technology Guwahati belonging to the age group of 20-40 years. These subjects represented different ethnic orientations and were chosen to have good ethnic mix. During the recording, the subject and the facilitator sat face to face across a table. A snap shot of the recording setup is shown in the Fig 1. The subject wore the headset microphone connected to the tablet-PC and the other devices were kept in front of the subject on the table for recording. In this phase of data collection the mobile phones were used in off-line mode, where the data is not transmitted through the wireless channel. Initially the subject contributed reading style speech data in English for 3-5 minutes duration. This was followed by two recordings of speech data of 6-8 minutes duration each in conversation style in English and in subject's favorite language, which happened to be his/her mother tongue in most cases. These languages are listed in Table II with their number of occurrences. In conversation style data collection, the subjects were prompted by the facilitators. Each subject also contributed the second session data after a gap of at least one week.

B. IITG-MV Phase-II Data set

Unlike in the Phase-I, the Phase-II data collection was done in uncontrolled environments such as laboratories, hostel rooms and corridors etc. This dataset contains speech from 70 male and 30 female subjects. For collecting the data, the same hardware setup as in Phase-I was employed with some differences in operating conditions. For the data recording in this phase, the facilitator called the subject on his/her own mobile phone from a distant place and the speech data was recorded in the mobile phone at the facilitator's end. Another mobile phone operating in off-line mode with hands-free microphone attached to the subject at the waist level also

Figure 1: A snapshot of recording setup of the IITG-MV Phase-I data collection in the office environment.

Table II: The list of the languages and their occurrences in the IITG-MV database under subjects' favorite language category

Language	Occurrence	
	Phase-I	Phase-II
Hindi	28	33
Telugu	10	21
Malayalam	15	8
Oriya	12	11
Bengali	4	9
Assamese	9	6
Gujarati	2	1
Tamil	8	4
Kannada	7	5
Nepali	1	1
Mizo	1	-
Marathi	2	-
English	1	1

recorded the speech data similar to that in Phase-I. A snap shot of a recording in a laboratory is shown in Fig 2. The other devices, Tablet PC and DVR were used in the similar fashion as done in Phase-I.

Figure 2: A snapshot of data recording in a laboratory for the IITG-MV Phase-II data collection.

III. Experimental Setup and Baseline System Performance

For the experiments a speaker verification system was developed using Gaussian mixture model with universal background model (GMM-UBM) based speaker modeling approach [1]. The configuration of the system is similar to that of simple

SV system fielded in the NIST 2003 speaker recognition evaluations (NIST SRE-03) [9].

A. System structure

The UBM was built with a mixture of 1024 Gaussian components with diagonal covariance matrices. 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. All speech data used is sampled at 8 kHz with 16 bits/sample resolution and was analyzed using a Hamming window of length 20 ms, frame rate of 100 Hz and pre-emphasis factor of 0.97. The MFCC feature vectors of 39 dimension were used to parameterize the speech data. Each feature vector comprised of C_1 to C_{13} static MFCC and their first and second order derivatives. To remove the non-speech portions from input data, an energy based voice activity detector with fixed threshold was used. The cepstral mean subtraction was applied on all features so as to reduce the effect of mismatch in channel. For finding out the performance of the SV system, the detection error trade-off (DET)curves were plotted using log likelihood ratio between claimant model and UBM. The equal error rate (EER) noted from the DET curve is used to quantify the speaker verification performance in all conditions. The SV system was developed using the hidden Markov model toolkit (HTK) [10].

B. Performance with NIST SRE-03 database

NIST SRE-03 database was used to benchmark the performance of the chosen structure of the SV system. It contains speech data of 356 target speakers (144 males and 212 female) collected over cellular phone network. Each speech file is a part of conversation between two speakers. The speech data is sampled at 8 kHz with 16 bits/sample resolution. The training set contains speech data of 2 minutes duration per speaker. The test set contains more than 3000 segments of varying length. The evaluation of the system is done as per the NIST SRE-03 evaluation plan for primary task [9], where all the test segments of length falling between 15-45 seconds have to be tested against the models specified. Each test segment is tested against 11 models out of which one may be a true trial and rest are false trials. The UBM was trained using 10 hours of speech data (balanced in gender) taken from Switchboard Cellular Part 2 corpus. The above explained experimental setup results in 24981 trials for verification task including true and false trials.

The performance of the SV system trained and tested on NIST SRE-03 database in terms of DET curves is shown in Figure 3. It is noted that the EER is 10.5 % which is close to that of the similar complexity system reported in NIST SRE-03 evaluation [11].

C. Performance with IITG-MV database

For evaluating the performance with our collected data, we have used the same SV system structure including the features as described above. To keep the recording condition same as that of the NIST SRE-03 database we have used the set of data

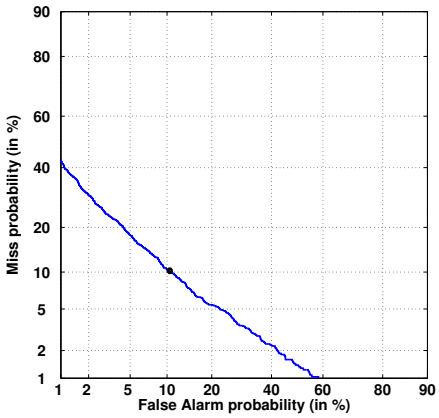

Figure 3: DET curve showing performance of the speaker verification system for NIST SRE-03 database

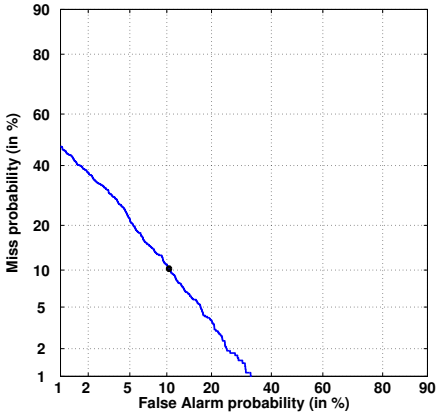

Figure 4: DET curve showing performance of the speaker verification system for IITG-MV Phase-I mobile phone data

collected over mobile phone condition from IITG-MV Phase-I data set. The training data set contains 100 speech segments of 2 minutes each from the first session recordings. The test set contains 1000 segments of speech, derived from the second session recording of the same data set with 10 segments for each speaker. The segments are of length varying from 30 to 45 seconds. Similar to NIST SRE-03 protocol, each test segment is tested against 11 models out of which one is a true trial. This makes a total of 11,000 test trials with a true trial to false trial ratio equal to 0.1. For building the UBM, we have used 10 hours of speech data from 50 speakers from IITG-MV Phase-II data set which are who are not common with speakers of Phase-I data set.

The performance in terms of DET curves for the SV system trained and tested on IITG-MV Phase-I data set from mobile phone condition is shown in Figure 3. It is noted that the EER is 10.3 % which is similar that observed for NIST SRE-03 database.

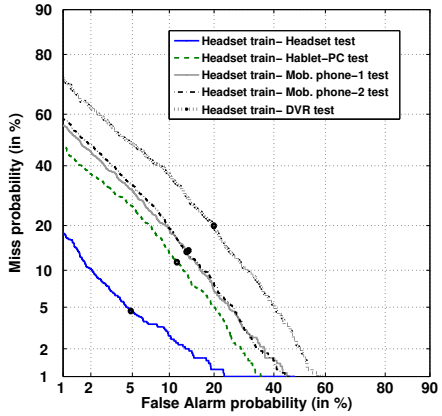

Figure 5: DET curves showing the performance of the SV system for sensor match and mismatch cases for IITG-MV Phase-I data.

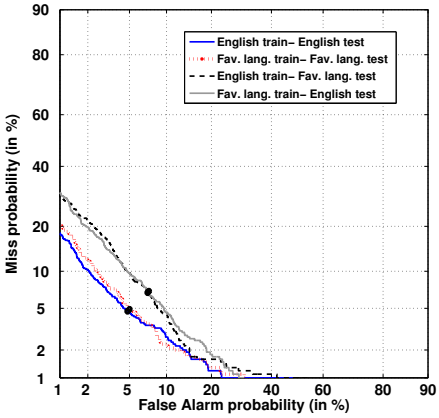

Figure 6: DET curves showing the performance of the SV system for language match and mismatch cases for IITG-MV Phase-I data.

IV. Study of Different Mismatches on Speaker Verification System Performance

To study how the mismatches in sensor, language, style of speech and environment between the training and testing affect the performance of a SV system, we have made use of the data collected under varying conditions in IITG-MV database. For these studies we have kept the SV system configuration same as described earlier.

A. Sensor mismatch

The IITG-MV Phase-I database contains speech data from subjects, recorded in parallel using five different sensors, namely, headset microphone, mobile phone-1, mobile phone-2, Tablet PC built-in microphone and digital voice recorder. To study the effect of these sensors on the performance of the system, the language, the style of speech and the environment of recording, were kept fixed across the training and testing data. For this experiment we have used following conditions: environment as office, language as English, style of speech as conversation. As the data collected with headset microphone was the cleanest and of higher SNR, we have used that for training the speaker models while the testing is done using data collected using all the five sensors. The performance for different sensors in terms of DET curves are shown in Figure 5. It is noted that the best EER of 4.7% is observed for the matched case as expected. In mismatch conditions, EER of 11% , 13.2% , 13.5% and 19.7% are observed for tablet-PC, mobile phone-1, mobile phone-2 and DVR, respectively. Thus the mismatch in training and testing sensors results in large degradation in performance.

B. Language mismatch

IITG-MV Phase-I data set contains speech data from 100 speakers in two languages: English and the subject's favorite language. The performance of the separate systems trained with English language data and with subject's favorite language data for matched and mismatched language testing conditions are given in terms of DET curves in Figure 6. It can be observed that in matched language cases, the performance of the system is similar irrespective of the language used for training and testing. Whereas, in case of mismatched language cases, the performances degrade by 2.5% in EER with respect to their matched case performances.

C. Speaking style mismatch

The conversational and reading style of speech have considerable difference and this is a well known fact. A typical SV system application such as secure access, may collect the enrollment data for a speaker in reading style, while the test data is more likely to be in conversational style. So, it would be interesting to study the impact of speaking style on SV system performance. To study this, we have used the 100 speakers data from IITG-MV Phase-I data set available in both conversational and reading styles while keeping the other conditions of the data as headset sensor, English language and office environment. The performances of the separate systems trained using conversational and reading data for matched and mismatched testing conditions are given in terms of DET curves in Figure 7. It is noted that, for mismatched testing cases a performance degradation of about 4% in EER compared to that of the matched case.

D. Environmental mismatch

To study the impact of environmental mismatch on the SV system performance, the data collected in office and in uncontrolled environments in IITG-MV Phase-I and Phase-II, respectively, was used. As across Phase-I and Phase-II there were 50 common speakers, the study was done on those set of 50 speakers rather than on 100 speakers set used for the above three studies. The required data for training and testing purposes, were created out of the 50 speakers set following the NIST SRE-03 evaluation plan for primary task as done for the earlier used 100 speakers data set. The other conditions of the data was kept as headset sensor, English language and reading style. The performances of the systems trained on office environment data when tested on office and

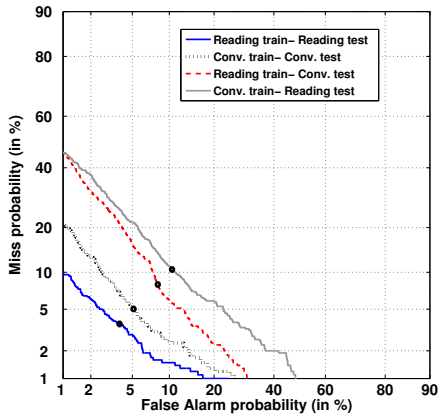

Figure 7: DET curves showing the performance of the system for reading style match and mismatch cases for IITG-MV Phase-I data.

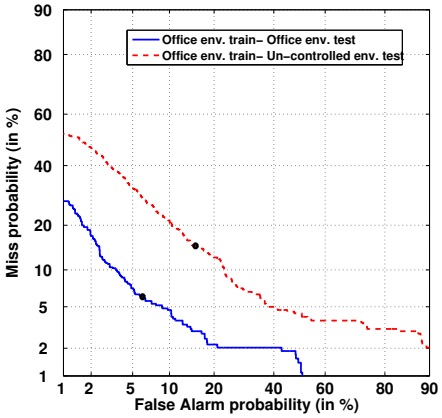

Figure 8: DET curves showing the performance of the system for the environmental match and mismatch cases for IITG-MV database.

uncontrolled environments data are given in terms of DET curves in Figure 8. It is to note that although the performance of matched case slightly degrades due to small test set size there is a degradation of about 7% in EER in notice in mismatched environment case compared to that of the matched case.

The comparable performance of the SV system developed using NIST SRE-03 and IITG-MV Phase-I databases under similar conditions provides the necessary assurance that our collected database can be reliably used for speaker recognition research purposes. The developed database reflects commonly used sensors, speaking styles and environment conditions in Indian context. The database also contains parallel recording in variety of conditions which includes mobile phone based speech collected with and without channel variations. The study conducted to explore effect of mismatch conditions on the SV system performance quantifies the impact of each of the sources of the mismatch.

V. Conclusion

A recently collected speech database for purpose of developing a robust speaker recognition system in Indian context is reported. Our collected database is found to give comparable speaker verification performance to that obtained using standard NIST SRE-03 database. It differs from other available public domain databases in containing parallel recording over different sensors and other variability to enable the assessment of their impacts and to promote more research on their modeling or compensation. In future we will explore existing techniques of mismatch reduction on IITG-MV database.

Our study exploring the impact of mismatch in training and test conditions with collected data finds that for the mismatch in sensor, in speaking style, and in environment results in significants degradation in performance compared to matched case whereas for mismatch in language case the degradation is found to be relatively smaller. Future effort should be in the direction of improving performance of speaker verification system to provide robustness against mismatch conditions in Indian scenario.

Acknowledgment

This work has been supported by the ongoing project grant No. 12(4)/2009-ESD sponsored by the Department of Information Technology, Government of India.

References

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[2] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, pp. 308–311, 2006.

[3] A. Adami, R. Mihaescu, D. Reynolds, and J. Godfrey, "Modeling prosodic dynamics for speaker recognition," in *Proc. ICASSP '03*, vol. 4, 2003, pp. IV–788–91 vol.4.

[4] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42–54, 2000.

[5] D. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. ICASSP '03*, vol. 2, 2003, pp. II–53–6 vol.2.

[6] R. Teunen, B. Shahshahani, and L. Heck, "A model-based transformational approach to robust speaker recognition," in *Proc. Int. Conf. on Spoken Language Processing (ICSLP '00)*, vol. 2, 2000, pp. 495–498.

[7] S.-C. Yin, R. Rose, and P. Kenny, "A joint factor analysis approach to progressive model adaptation in text-independent speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 1999–2010, 2007.

[8] A. Solomonoff, W. Campbell, and I. Boardman, "Advances In Channel Compensation For SVM Speaker Recognition," in *Proc. ICASSP '05*, 18 2005.

[9] NIST 2003 Speaker Recognition Evaluation Plan, http://www.itl.nist.gov/iad/mig/tests/sre/2003/2003-spkrec-evalplan-v2.2.pdf.

[10] S. Young, G. Evermann, M. Gales, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book version 3.4*. Cambridge, U.K.: Cambridge University Engineering Department, 2006.

[11] M. P. Alvin and A. Martin, "Nist speaker recognition evaluation chronicles," in *Proc. Odyssey 2004, The Speaker and Language Recognition Workshop*, 2004, pp. 12–22.