Multivariability Speaker Recognition Database in Indian Scenario

Haris B C, G. Pradhan, A. Misra, S. R. M. Prasanna, R. K. Das and R. Sinha

Department of Electronics and Communication Engineering,
Indian Institute of Technology Guwahati, Guwahati-781039, India
{*haris, gayadhar,abhinavmisra, prasanna, rohankd, rsinha*}*@iitg.ernet.in*

Abstract

In this paper we describe the collection and organization of the speaker recognition database in Indian scenario named as IITG Multivariability Speaker Recognition Database. The database contains speech from 451 speakers speaking English and other Indian languages both in conversational and read speech styles recorded using various sensors in parallel under different environmental conditions. The database is organized into four phases on the basis of different conditions employed for the recording. The results of the initial studies conducted on a speaker verification system exploring the impact of mismatch in training and test conditions using the collected data are also included. A copy of this database can be obtained from the authors by contacting them.

Key words: multivariability, Indian scenario, speaker recognition, speaker verification.

1 Introduction

The availability of standard speaker recognition databases helped in benchmarking of the speaker recognition [Doddington, 1985, Reynolds, 2002] systems at different places. The availability of the databases may be one factor to the progress made in the speaker recognition field [Campbell and Reynolds, 1999, Reynolds, 2002]. In developing and evaluating speaker verification system, the standard database has an important role. It facilitates the comparison of performances of different techniques on common data, thus enabling us to know which of them is more promising to pursue. It also helps in determining the state-of-the art performance under particular experimental conditions and highlights the issues that require further exploration. Till now, mostly American English has been the focal point of the speaker recognition research. Comparatively much less work has been done to document the speech and speaker characteristics in Indian languages and scenario [Patil and Basu, 2008, Patil et al., 2006]. Also, the recent advances in speaker verification research such as the joint factor analysis (JFA) [Yin et al., 2007] and the total variability *i-vector* [Dehak et al., 2011] based systems require data recorded with various conditions like recording environment and sensors for each speaker, for building the system. Through this effort we intend to fill this gap by building a multivariability database with parallel recordings in Indian scenario. We have tried to make the newly created database as multipurpose as possible. It can be used to conduct multiple speech related studies like the effect of noise and reverberation, sensor mismatch, language variation, style variation, narrowband and wideband signal analysis, effect of coding in speaker recognition, speaker diarization etc. The various criteria for maintaining the quality of the database [Patil and Basu, 2008] have been taken care while collecting the data.

The database were collected keeping in view the Indian scenario where there is a wide variability present in the languages, styles, environmental conditions and sensors. The entire corpus aims to aid automatic speaker recognition research in the

direction of building systems that can perform equally well in the changing recording conditions. Study of sensor, language, style and environmental variations can be done in the Indian context. While taking recordings, the placement of the sensors was such that speech data is collected in degraded conditions like background noise and reverberation. To impart a practical dimension to the corpus, the choice of the sensors was such that they are portable, low cost and used extensively in public domain. The speakers were carefully selected to represent distinct parts of India speaking different languages. They were all from different educational backgrounds, ethnic orientations and age. All this ensured the needed inter-speaker variability. In order to get the intra-speaker variability, two sessions were taken for each speaker in which they read the same passages, but may be conversing in different way.

The database is created aiming to support and evaluate the automatic speaker recognition systems where channel, language, style and environments may vary. Accordingly, the database is named as IIT Guwahati (IITG) multi-variability (MV) speaker recognition database [Haris B C et al., 2011]. It contains four sets, namely, IITG MV Phase-I, Phase-II, Phase-III and Phase-IV. The IITG-MV Phase-I dataset is collected from 100 subjects over two sessions in an office environment involving multiple sensors, multiple languages, and different speaking styles (conversational and read speech). IITG MV Phase-II also contains data of 100 speakers and differs from the Phase-I by collecting speech data mainly in multiple environments, namely, laboratories and hostel rooms, while keeping the other variabilities unchanged. In the third phase of recording, a truly conversational style telephonic speech data is collected and it is termed as IITG MV Phase-III dataset. Finally, in Phase-IV, with the view of supporting the development of a remote person authentication system, speech data was collected from the speakers all over the country through an interactive voice response (IVR) System and is termed as IITG MV Phase-IV dataset. The Phase IV of speech data is collected in three parts. The Part-I of Phase-IV involved speech data collection from 55 subjects all over India and to be used for the development of universal background model (UBM). The Part-II of Phase-IV involved speech data collection from 89 subjects all over India and to be used for the development of speaker models. The Part-III of Phase-IV has speech from 197 genuine trails and 130 imposter trails.

The rest of the report is organized as follows: Section 2, Section 3, Section 4 and Section 5 describes the development and organization of Phase I, Phase II, Phase III and Phase IV databases, respectively. Section 6 explains the initial speaker verification studies conducted using the collected database and discusses the results obtained. Section 7 provides the summary of the work and possible scope for the future work.

2 IITG MV Phase-I speaker recognition database

IITG MV Phase-I database contains two sessions speech data from 100 speakers collected in an office environment. In this database the following three types of variabilities were considered while collecting the speech data from different speakers:

- Multi-sensor: Speech data were recorded over five different sensors simultaneously.
- Multi-lingual: Every speaker spoke in two different languages, namely, English and his/her favorite Indian language.
- Multi-style: Every speaker spoke in reading and conversational styles.

2.1 Speech data collection

One of the faculty chambers (measuring $15 \times 12 \times 12$ feet) was used as an office environment. Apart from the air conditioner and a fan no other electrical equipments were working in the room. Both the windows and doors were closed during recording, to avoid any external noise coming into the room. The idea was to capture the effects of reverberation and ambient noise in the room effectively. The following were the recording sensors used:

- Headset Microphone: The headset microphone that comes with most personal computers (PCs) was used as the sensor. It has wideband characteristics with a flat response up to 16 kHz, thus enabling us to collect speech data at 16 kHz sampling

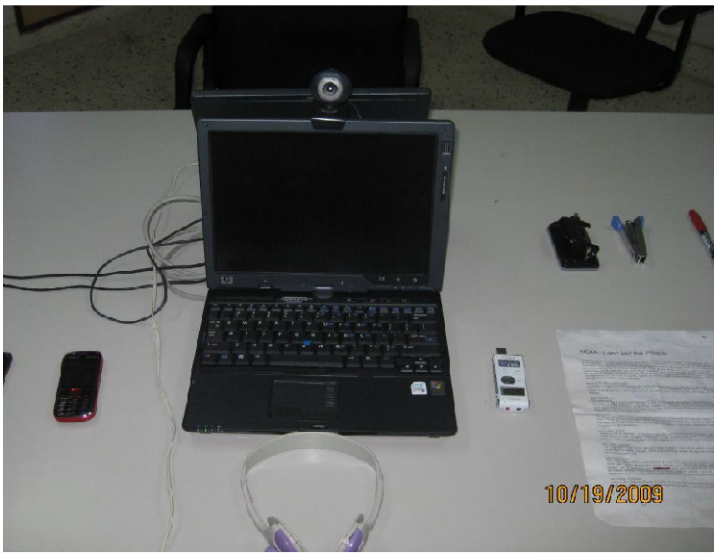

Fig. 1. Placement of sensors for Phase-I data recording

frequency with 16 bits per sample resolution. The headset microphone is an omni-directional one. The headphone was mounted close to the speaker, so that the best possible clean speech data is obtained compared to all other sensors.

- In-built microphone: The in-built microphone that comes with Tablet PCs was used as another sensor. The in-built microphone is also of wideband characteristics and also omni-directional in nature. The speech data is also collected using a sampling frequency of 16 kHz with 16 bits per sample resolution.
- Digital Voice Recorder: A digital voice recorder was used for recording the data. This included a sensor that has wideband characteristics of at least 22 kHz and omni-directional in nature. It was the most sensitive of all the sensors and recorded in stereo (two) channel, MP3 format with sampling frequency of 44 kHz and 16 bits per sample resolution. After transferring data from the recorder to the PC, the MP3 files were converted to WAV format with the sampling frequency of 16 kHz, mono channel and 16 bits per sample resolution.
- Mobile phones in offline mode: Two mobile phones of different make were used for recording the data. Both the phones have a voice recorder software that stores files in AMR format at a sampling frequency of 8 kHz and stored with 16 bits per sample resolution and were placed on the table at a distance of 2-3 feet from the subject. The AMR files were converted into WAV format after storing them in the PC.

The technical details of the sensors used for recording are summarized in the Table 1. Figure 1 shows the placement of sensors and Figure 2 shows the way recording was done in the office environment. The subjects for recording included members of the student, staff and faculty community from IIT Guwahati in the age group of 20-40. Speech data for about 3-5 minutes in reading style using English passage was initially collected. This was followed by speech data for about 6-8 minutes of two recordings in conversation style for both English and favorite language, where later happens to be the mother tongue in most cases. During the entire recording a facilitator was present to direct the subject and also to converse with him for recording. The second session of recording for each speaker was done after a gap of around one week. The languages which fall in the category of favorite language are listed in Table 2 with their number of occurrences.

Fig. 2. Snapshot depicting recording scenario in case of Phase-I data collection.

Table 1
Technical details of the sensors/devices used for collecting the speech data

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

2.2 Processing of the data

After recording, data from each sensor was processed using a computer to remove facilitator's voice in a semi-automatic fashion. The Wavesurfer [www.speech.kth.se/wavesurfer/, -] software was used to open a subject's speech file and mark the start and end points of the regions where facilitator is instructing or asking questions. These points are then used as inputs to a MATLAB code that removes the facilitator's voice. The speech data from all sensors is stored in uncompressed WAV file format with 16-bit resolution.

2.3 Nomenclature for speech files

After processing, all of the speech files were named with a unique identification code which contains fields corresponding to speaker's ID , session, sensor, environment, language and style. The details of the identification code are given below:

- The file name is of THIRTEEN (13) characters, where first FOUR(4) digits represent the Unique speaker ID, and the remaining FIVE(8) alphabets followed by underscore(_) represents the remaining recording condition variables.
- Sequence of naming is : <Unique speaker ID> <Underscore> <Session ID> <Recording sensor ID> <Environment ID> <Language ID> <Style ID>

Table 2
The list of the languages and their occurrences in the IITG MV Phase-I dataset under subjects' favorite language category

Language	Occurrence
Hindi	28
Telugu	10
Malayalam	15
Oriya	12
Bengali	4
Assamese	9
Gujarati	2
Tamil	8
Kannada	7
Nepali	1
Mizo	1
Marathi	2
English	1

- Speaker ID: This is the unique ID representing the subject using four decimal digits. Male speakers IDs start with '1' and that of female speakers start with '2'.
- Session ID: This field shows the session of the recording by a single letter. Letter 'A' represents the first session of recording and letter 'B' represents the second session.
- Language ID : This field shows the language of the recording by two letters. Table 4 gives the code for all the languages present in the database.
- Sensor ID : This field shows the sensor used for recording the speech by three alphanumeric characters. H01 = Headset (HCL); T01 = Tablet PC in front of subject; M01 = Mobile Phone (Nokia 5130c); M02 = Mobile Phone (Sony Ericsson W350i); D01 = Digital Voice Recorder (Sony)
- Environment ID: This field shows the environment of recording. In IITG MV Phase-I each speaker's data is collected in the office environment which is represented by the letter 'O'.
- Style ID: Single capital letter representing the style of recorded speech. 'R' represents reading style and 'C' represents conversational style.
- The details about different fields in the unique identification code is summarized in 3. The additional details about subjects viz. age, language spoken frequently at home and language of instruction during schooling are summarized in the Table 5.

Example: 1001_BM02OKNC.wav means Speaker is a male with ID 1001, second session, sensor used is mobile phone (Sony Ericsson), recording is done in office environment and the language is 'Kannada' in conversational style.

2.4 Organization of the database

After giving appropriate names, each of the recorded speech files were stored in a structured fashion. Each session data consists of THREE sub-sessions:

(1) First sub-session data consists of read speech in English language recorded using five sensors. Each speech file is of duration 3 minutes on average.
(2) Second sub-session data consists of conversational speech in English language recorded using five sensors. Each speech file is of duration 3 minutes on average.

Table 3
Nomenclature for speech file names

Variable (code)	Description	Nomenclature
Speaker ID	Unique ID representing the subject	Four decimal digits; male speakers IDs start with '1' and female speakers IDs start with '2'
Session ID	Mentions the session number of the recording	Single capital letter (A = 1st Session .. Z = 26th Session)
Sensor ID	Mentions the recording sensor used	Three characters. First, a single char (T=Tablet PC; D=Digital Voice Recorder; H=Headset M=Mobile Phone(s)). Next, two numbers (01,02,..)
Environment ID	Mentions the recording environment	Single capital letter (O=Office Environment, M=Multi Environment)
Language ID	Mentions language of speech recorded	Two capital letters. Ref. Table 4 for details.
Style ID	Mentions style of recorded speech	Single capital letter (R=Reading Text; C=Conversation)

Table 4
Table showing Language ID for different languages

Language	Language ID	Language	Language ID
Assamese	AS	Manipuri	MN
Bengali	BN	Marathi	MR
Bodo	BD	Nepali	NP
Dogri	DG	Oriya	OR
English	EN	Punjabi	PN
Gujarati	GJ	Sanskrit	SN
Hindi	HN	Santali	ST
Kannada	KN	Sindhi	SD
Kashmiri	KS	Tamil	TM
Konkani	KK	Telugu	TL
Maithili	MT	Urdu	UR
Malayalam	ML	Mizo	MZ

(3) Third sub-session data consists of conversational speech in speakers favorite language recorded using five sensors. Each speech file is of duration 3 minutes on average.

The complete database is stored in the following directory structure:

Speech_Data → Office_Environment → Session1(Session2) → English_Language(Favorite_Language)
→Sensors(D01,H01,M01,M02,T01)→SpeechFiles.

3 IITG MV Phase-II speaker recognition database

Unlike in the Phase-I, the Phase-II data collection was done in uncontrolled environments such as laboratories, hostel rooms and corridors etc. Other differences of the Phase-II database compared to Phase-I are that it contains speech data passed through wireless mobile channel and a read/conversational speech recorded in Hindi language from all the speakers. The variabilities present in the Phase-II database are four as listed below.

Table 5
Table showing details about the speakers in IITG Phase-I database

Speaker ID	Gender (M/F)	Age	Language spoken frequently at home	Language of instruction during schooling	Speaker ID	Gender (M/F)	Age	Language spoken frequently at home	Language of instruction during schooling
1001	M	38	KN	EN	1002	M	28	HN	EN
1005	M	24	HN	HN	1011	M	26	ML	ML
1012	M	24	ML	EN	1013	M	22	ML	ML
1015	M	24	TL	EN	1016	M	26	TL	TL
1017	M	22	TL	EN	1018	M	21	KN	EN
1019	M	21	ML	EN	1020	M	23	TL	EN
1021	M	23	BN	BN	1022	M	24	HN	EN
1024	M	22	TL	TL	1025	M	23	TL	TL
1026	M	31	ML	EN	1027	M	25	ML	ML
1028	M	26	ML	EN	1031	M	24	TL	TL
1032	M	24	ML	ML	1036	M	21	HN	HN
1037	M	30	AS	EN	1038	M	39	OR	OR
1039	M	23	MZ	EN	1042	M	22	HN	EN
1043	M	24	HN	HN	1044	M	22	HN	EN
1045	M	27	TM	TM	1049	M	21	HN	EN
1051	M	22	BN	EN	1052	M	27	TM	EN
1053	M	23	KN	EN	1054	M	24	HN	EN
1055	M	26	BN	EN	1057	M	27	HN	HN
1058	M	39	AS	AS	1061	M	30	GJ	GJ
1062	M	23	HN	EN	1063	M	24	MR	MR
1064	M	29	OR	OR	1065	M	25	HN	HN
1066	M	24	HN	HN	1067	M	23	OR	OR
1068	M	26	OR	OR	1069	M	31	OR	OR
1070	M	32	OR	OR	1071	M	28	ML	ML
1072	M	33	GJ	GJ	1073	M	30	HN	EN
1074	M	34	OR	OR	1075	M	25	HN	HN
1076	M	28	ML	ML	1077	M	31	OR	OR
1078	M	38	KN	KN	1079	M	31	OR	OR
1080	M	32	TM	EN	1081	M	28	HN	EN
1082	M	37	AS	AS	1084	M	32	OR	OR
1085	M	24	BN	BN	1086	M	23	HN	HN
1087	M	20	HN	EN	1088	M	25	HN	HN
1089	M	20	HN	HN	1090	M	26	MR	EN
1091	M	26	HN	EN	1092	M	28	TL	EN
1093	M	24	HN	HN	1094	M	30	KN	EN
1095	M	25	OR	EN	1096	M	29	AS	EN
1098	M	28	TM	TM	1099	M	30	TM	EN
1101	M	29	HN	HN	1104	M	27	TL	TL
1105	M	23	TL	TL	1106	M	29	AS	EN
1107	M	22	HN	EN	1108	M	32	OR	OR
1109	M	26	HN	EN	2014	F	26	TM	EN
2023	F	25	ML	ML	2029	F	28	ML	ML
2030	F	27	ML	ML	2033	F	24	ML	ML
2034	F	20	EN	EN	2035	F	21	ML	ML
2040	F	26	KN	EN	2041	F	40	KN	KN
2046	F	23	HN	EN	2047	F	26	HN	HN
2048	F	25	AS	AS	2050	F	22	TM	EN
2056	F	39	AS	EN	2083	F	27	TM	EN
2097	F	26	NP	EN	2100	F	29	AS	EN
2102	F	28	AS	EN	2103	F	24	HN	EN

- Multi-environment: Speech data were recorded in different environments like hostel rooms and laboratories.
- Multi-sensor: Speech data were recorded over five different sensors simultaneously.
- Multi-lingual: Every speaker spoke in three different languages, namely, English, Hindi and his/her favorite language.
- Multi-style: Every speaker spoke in reading and conversational styles.

3.1 Speech data collection

Speech data was collected mainly from two kinds of environments, namely, laboratory and hostel rooms. For ensuring a practical scenario the recordings were done while people were working in the laboratory and computers, air-conditioners and other electrical equipments were on. At times it used to happen that in midst of recording, people would enter or leave lab, thereby adding to the ambient noise. A typical laboratory chamber was of around $25 \times 30 \times 15$ feet. Figure 3 shows the snap shot of speech data recording in a lab environment. Hostel rooms, on the other hand, were small, of around $10 \times 10 \times 15$ feet. At the time of recording only subject was sitting inside the small room. Usually windows were open and doors were half closed. Noise from people walking and talking in the corridors could be heard easily. Other than the electric fan, no other electrical equipments were working in the room at the time of recording. A snap shot of a recording in a hostel room is shown in Fig 4.

Fig. 3. Snapshot of recording inside laboratory for Phase II data collection

For collecting the speech 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 offline mode with hands-free microphone attached to the subject at the waist level also was used to record the speech data. The other devices, Tablet PC and DVR were used in the similar fashion as done in Phase-I. The arrangement of devices for recording speech in Phase-II data collection is shown in Fig 5. Refer Table 1 for the technical details of the sensors used for recording the speech data. The subjects for recording included members of the student, staff and faculty community from IIT Guwahati in the age group of 20-40. Read speech data for about 3-5 minutes in English language was initially collected. This was followed by a recording of speech data for about 6-8 minutes in conversation style for both English and favorite language, where later happens to be the mother tongue of the subject in most cases. Finally a read or conversational speech in Hindi language was recorded according to the subject's choice. During the entire recording a facilitator was present to direct the subject and also to converse with him for recording. The second session of recording for each speaker was done after a gap of around one week. The languages which fall in the category of favorite language are listed in Table 6 with their number of occurrences.

Fig. 4. Snapshot of recording inside hostel room for Phase II data collection

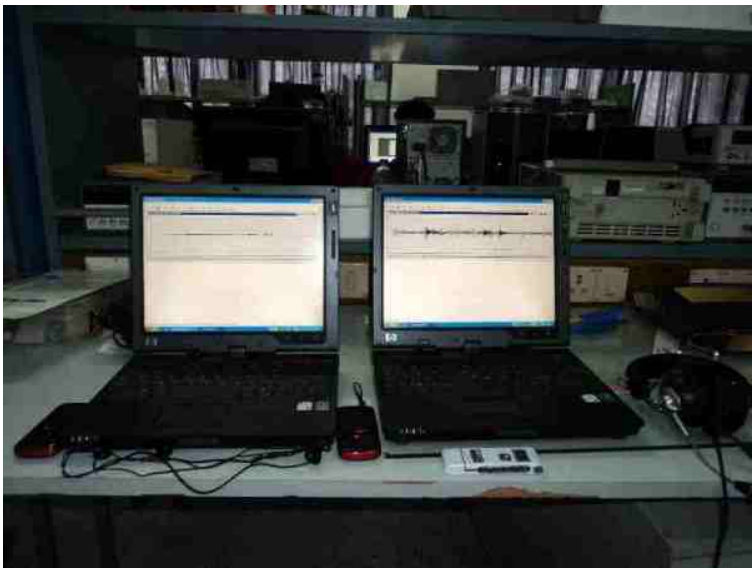

Fig. 5. Arrangement of sensors for Phase-II data collection

3.2 Processing of the data

After recording, data from each sensor was processed using a computer to remove facilitator's voice in a semi-automatic fashion. The Wavesurfer [www.speech.kth.se/wavesurfer/, -] software was used to open a subject's speech file and mark the start and end points of the regions where facilitator is instructing or asking questions. These points are then used as inputs to a MATLAB code that removes the facilitator's voice. The speech data from all sensors is stored in uncompressed WAV file format with 16-bit resolution.

3.3 Nomenclature for speech files

After processing, all of the speech files were named with a unique identification code which contains fields corresponding to speaker's ID , session, sensor, environment, language and style. The nomenclature of speech files follow the same as that of

Table 6
The list of the languages and their occurrences in the IITG MV Phase-II dataset under subjects' favorite language category

Language	**Occurrence**
Hindi	33
Telugu	21
Malayalam	8
Oriya	11
Bengali	9
Assamese	6
Gujarati	1
Tamil	4
Kannada	5
Nepali	1
English	1

Phase-I database except for the following cases:

- Sensor ID : This field shows the sensor used for recording the speech by three alphanumeric characters. H01 = Headset (HCL); T01 = Tablet PC in front of subject; M03 = Mobile Phone (Nokia 5130c)- online recording; M04 = Mobile Phone (Nokia 5130c)- offline recording; D01 = Digital Voice Recorder (Sony)
- Environment ID: This field shows the environment of recording. In IITG MV Phase-II each speaker's data is collected in multi-environments which is represented by the letter 'M'.
- For details about deferent fields in the unique identification code, refer Table 3. The additional details about subjects viz. age, language spoken frequently at home , language of instruction during schooling and place of recording are summarized in the Table 7.

3.4 Organization of the database

After giving appropriate names, each of the recorded speech files were stored in a structured fashion. Each session data consists of FOUR sub-sessions:

(1) First sub-session data consists of read speech in English language recorded using five sensors. Each speech file is of duration 3 minutes on average.
(2) Second sub-session data consists of conversational speech in English language recorded using five sensors. Each speech file is of duration 3 minutes on average.
(3) Third sub-session data consists of conversational speech in speakers favorite language recorded using five sensors. Each speech file is of duration 3 minutes on average.
(4) Third sub-session data consists of read/conversational speech in Hindi language recorded using five sensors. Each speech file is of duration 3 minutes on average.

The complete database is stored in the following directory structure:

Speech_Data → Multi_Environment → Session1(Session2) → English_Language(Favorite_Language, Hindi_language)
→ Sensors(D01,H01,M01,M02,T01) → SpeechFiles.

Table 7
Table showing details about the speakers in IITG Phase-II database

Speaker ID	Gender (M/F)	Age	Language spoken frequently at home	Language of instruction during schooling	Place of recording (L-lab, R- hostel room)		Speaker ID	Gender (M/F)	Age	Language spoken frequently at home	Language of instruction during schooling	Place of recording (L-lab, R- hostel room)	
					session-1	session-2						session-1	session-2
1001	M	38	KN	EN	L	L	1005	M	24	HN	HN	L	L
1011	M	26	ML	ML	R	R	1013	M	22	ML	ML	L	L
1016	M	26	TL	TL	R	L	1018	M	21	KN	EN	L	L
1020	M	23	TL	EN	L	R	1026	M	31	ML	EN	R	L
1028	M	26	ML	EN	L	R	1038	M	39	OR	OR	L	L
1043	M	25	HN	HN	R	R	1044	M	22	HN	EN	R	R
1051	M	22	BN	EN	R	R	1058	M	39	AS	AS	L	L
1061	M	30	GJ	GJ	R	R	1062	M	23	HN	EN	R	R
1063	M	24	MR	MR	R	R	1064	M	29	OR	OR	R	R
1065	M	25	HN	HN	R	R	1066	M	24	HN	HN	R	R
1067	M	23	OR	OR	R	R	1068	M	26	OR	OR	R	L
1069	M	31	OR	OR	L	L	1070	M	32	OR	OR	L	L
1071	M	28	ML	ML	L	L	1073	M	30	HN	EN	L	L
1074	M	34	OR	OR	L	L	1075	M	25	HN	HN	R	R
1076	M	28	ML	ML	L	L	1077	M	31	OR	OR	R	R
1080	M	32	TM	EN	L	L	1081	M	28	HN	EN	R	R
1085	M	24	BN	BN	R	R	1089	M	21	HN	HN	R	R
1093	M	24	HN	HN	R	R	1094	M	30	KN	EN	L	L
1101	M	29	HN	HN	L	L	1104	M	27	TL	TL	L	L
1108	M	32	OR	OR	R	L	1141	M	25	OR	OR	R	R
1142	M	24	MR	EN	R	R	1144	M	25	AS	AS	R	R
1145	M	25	OR	OR	R	R	1146	M	26	HN	EN	R	R
1147	M	22	HN	HN	R	R	1148	M	24	BN	BN	R	R
1149	M	22	TL	TL	R	R	1150	M	23	TL	EN	R	R
1200	M	25	HN	EN	L	R	1201	M	21	BN	EN	L	R
1202	M	26	AS	EN	L	L	1207	M	21	TL	TL	R	L
1208	M	22	TL	EN	R	R	1209	M	21	HN	EN	R	L
1210	M	21	HN	HN	R	R	1211	M	22	TL	EN	R	L
1212	M	20	HN	EN	R	R	1252	M	23	BN	EN	L	L
1253	M	24	HN	EN	L	R	1255	M	23	HN	HN	L	L
1256	M	21	TL	EN	L	R	1257	M	21	TL	EN	R	R
1258	M	21	TL	EN	R	R	1259	M	21	TL	EN	R	R
1260	M	22	TL	EN	R	R	1261	M	23	TL	EN	R	L
1262	M	22	TL	EN	R	L	1263	M	20	TL	EN	R	R
1264	M	22	TL	EN	R	R	1265	M	22	TL	EN	R	R
2014	F	26	TM	EN	L	L	2033	F	21	ML	ML	R	R
2034	F	20	EN	EN	R	L	2040	F	26	KN	EN	R	R
2041	F	40	KN	KN	L	L	2046	F	23	HN	EN	L	L
2047	F	26	HN	HN	L	L	2048	F	25	AS	AS	R	R
2083	F	27	TM	EN	L	L	2097	F	26	NP	EN	L	L
2100	F	29	AS	EN	L	L	2102	F	28	AS	EN	L	L
2103	F	24	HN	EN	L	L	2110	F	20	HN	HN	R	R
2112	F	22	HN	EN	R	R	2113	F	23	HN	EN	R	R
2114	F	18	HN	HN	R	R	2115	F	18	TL	EN	R	L
2116	F	24	BN	EN	R	R	2117	F	22	BN	EN	R	R
2118	F	24	HN	EN	R	R	2119	F	21	HN	HN	R	R
2122	F	22	TM	EN	R	R	2123	F	23	HN	HN	R	R
2124	F	17	HN	EN	R	R	2125	F	18	HN	EN	R	R
2126	F	18	TL	EN	R	R	2127	F	23	HN	EN	R	R
2128	F	22	MR	EN	R	R	2129	F	22	HN	EN	R	R

4 IITG MV Phase-III speaker recognition database

In the Phase-III of recording the telephone network was used keeping in mind about the possible remote person authentication using speech mode. Unlike in phase I and phase II, in phase III a facilitator connects a call between two person in their free time in the conference call mode. The variabilities present in the Phase-III database are four as listed below.

- Multi-environment: Speech data were recorded by conversing in all kinds of practical environments possible like coffee shops, working places, rooms, Laboratories etc.
- Multi-sensor: Speech data were recorded over different mobile handsets at sampling frequency of 8 kHz.
- Multi-lingual: Every speaker spoke either in English or his/her mother tongue (favorite language).
- Conversation style: Every speaker spoke in conversational style over a conference call.

4.1 Speech data collection

Speech data was collected over conference call mode between two speakers when a facilitator connects the call in conference mode. The subjects were requested to engage in a natural conversation with any of their friends or relatives. In their whole conversation they mainly used two languages: their mother tongue and English. Predominantly, the subjects conversed in their mother tongue (favorite language), but in between they switched to English as well. While taking recordings, subjects were in different degraded environmental conditions that included background noise and reverberation as well. To impart a practical dimension to the corpus, we chose to record through the personal mobile handsets of the subjects themselves. This way we intended to cover all the different types of mobile sensors that are used by the public. Two sessions were taken for each speaker in which they conversed to the same person, but on different topics.

Figure 6 shows the way recording was done. The subjects were requested to give an appointment or a stipulated time at which they and their friend would be free to talk to each other. At the specified time, the facilitator, then called the subject from his mobile handset. The subject, in turn, gave the mobile number of the person he wants to talk to. The facilitator, subsequently put subject on hold, and called the other person. The two persons were then connected to each other by the facilitator, through a conference call. There was no precondition on the place of recording and the language of conversation. Subjects were engaged freely in conversation where language and matter of discussion changed frequently during the entire duration. For around fifteen minutes, the subjects talked to each other, while the call was being recorded at the facilitator's handset (Guwahati). In a very similar fashion, session two of the recording was taken for the same subject pair. On an average, the time gap between the two sessions was one week. This way, data was recorded from 100 speaker pairs leading to hundred conversations per session. Also, with a research perspective in mind, in almost 50 percent (47 out of 100) of the conversations, one of the subjects is common to both of the previously collected, Phase-I and Phase-II databases.

4.2 Processing of the data

After collecting this 100 speaker pairs conversational style database, we did post processing on it to create two different types of databases. The first is conversational style one speaker recognition database and the other is conversational style two speaker recognition database. The Wavesurfer [www.speech.kth.se/wavesurfer/, -] software was used to open a subject's speech file and mark the start and end points of the regions of the two different speakers. The speech data thus collected is stored in uncompressed WAV file format with 16-bit resolution.

4.3 Nomenclature for speech files

After processing, all of the speech files were named with a unique identification code which contains fields corresponding to speaker's ID , session, sensor, environment, language and style. The nomenclature of speech files of conversational style one

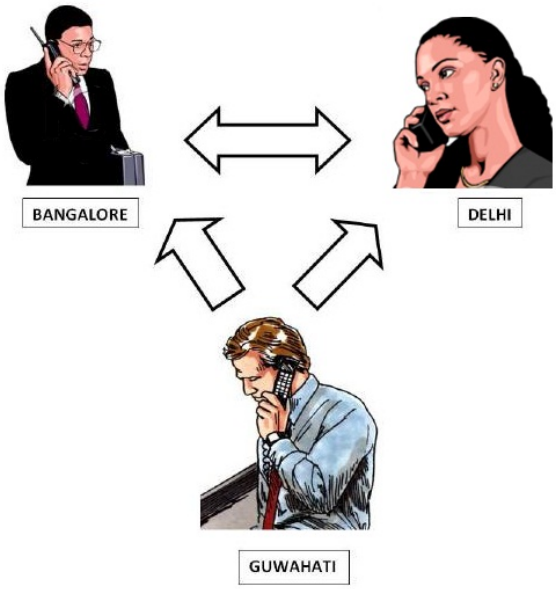

Fig. 6. Conference Call scenario for Phase III data collection

Table 8
The list of the languages and their occurrences in the IITG MV Phase-III dataset under subjects' favorite language category

Language	**Occurrence**
Hindi	62
Telugu	26
Malayalam	34
Oriya	24
Bengali	6
Assamese	18
Gujarati	2
Tamil	6
Kannada	8
Marathi	2
English	8
Marwari	2
Bhojpuri	2

speaker recognition database of Phase-III follow the same as that of Phase-I and Phase-II. But the nomenclature of speech files of conversational style two speaker recognition database of Phase III is different and as follows:

- The file name is of EIGHTEEN (18) characters, where first FOUR (4) digits represent the Unique Speaker ID of one speaker, following FOUR (4) digits after hyphen(-) represent the speaker ID of the other speaker and the remaining EIGHT (8) alphabets/numerals followed by underscore(_) represents the recording condition variables.
- Sequence of naming is : <Unique speaker ID of first speaker> <hyphen> <Unique speaker ID of second speaker> <Underscore> <Session ID> <Recording sensor ID> <Environment ID> <Language ID> <Style ID>

Table 9
Nomenclature for speech file names

Variable (code)	Description	Nomenclature
Speaker ID of first speaker	Unique ID representing the first subject	Four decimal digits; male speakers IDs start with '1' and female speakers IDs start with '2'
Speaker ID of second speaker	Unique ID representing the second subject	Four decimal digits; male speakers IDs start with '1' and female speakers IDs start with '2'
Session ID	Mentions the session number of the recording	Single capital letter (A = 1st Session .. Z = 26th Session)
Sensor ID	Mentions the recording sensor used	Three characters. First, a single char (M=Mobile Phone(s)). Next, two numbers (03)
Environment ID	Mentions the recording environment	Single capital letter (M=Multi Environment)
Language ID	Mentions language of speech recorded	Two capital letters. Ref. Table 4 for details.
Style ID	Mentions style of recorded speech	Single capital letter (C=Conversation)

- Speaker ID: This is the unique ID representing the subject using four decimal digits. Male speakers IDs start with '1' and that of female speakers start with '2'.
- Session ID: This field shows the session of the recording by a single letter. Letter 'A' represents the first session of recording and letter 'B' represents the second session.
- Language ID : This field shows the language of the recording by two letters. Table 4 gives the code for all the languages present in the database.
- Sensor ID : This field shows the sensor used for recording the speech by three alphanumeric characters. M03 = mobile phone.
- Environment ID: This field shows the environment of recording. In IITG MV Phase-III each speaker's data is collected in multi environment which is represented by the letter 'M'.
- Style ID: Single capital letter representing the style of recorded speech. 'C' represents conversational style.
- The details about different fields in the unique identification code is summarized in Table 9. The additional details about subjects viz. age, language spoken frequently at home and language of instruction during schooling are summarized in the Table 10.

Example: 1001-2001_BM03MKNC.wav means Speaker one is 1001, speaker two is 2001, Second Session, Mobile Phone (used by Facilitator to record voice), Mutli-Environment and Kannada language in Conversation mode is used for recording.

The database for Phase III of recording can be used for both one and two speaker recognition tasks. In the two speaker recognition task, the original conversation speech files are used without segmentation.

4.4 Organization of the database

After giving appropriate names, each of the recorded speech files were stored in a structured manner. The phase III database is divided in two parts, which are

- Part I: It contains 200 speakers conversational database (one speaker recognition database). The conversational data of 100 pairs of speaker is divided in 200 speaker single speaker recognition database, where the length of conversation is 15-20 minutes on average for both first and second session.

  The complete database is stored in following structure:

  IITG-MV Phase III Speaker Recognition Database Part I(single speaker recognition)→conversational style one speaker recognition→Session1(Session2)→Speech Files.

Table 10
Table showing details about the speakers in IITG Phase-III database

Speaker ID	Gender (M/F)	Age	Language spoken frequently at home	Speaker ID	Gender (M/F)	Age	Language spoken frequently at home	Speaker ID	Gender (M/F)	Age	Language spoken frequently at home
1001	M	40	KN	1005	M	25	HN	1011	M	27	ML
1012	M	25	ML	1013	M	24	ML	1018	M	23	KN
1019	M	23	ML	1026	M	35	ML	1028	M	27	ML
1038	M	39	OR	1043	M	25	HN	1044	M	25	TL
1051	M	25	BN	1058	M	40	AS	1061	M	40	GJ
1062	M	24	HN	1063	M	25	MR	1064	M	30	OR
1065	M	24	HN	1066	M	24	HN	1067	M	26	OR
1068	M	26	OR	1069	M	34	OR	1070	M	40	OR
1071	M	25	ML	1073	M	28	HN	1074	M	35	OR
1076	M	29	ML	1077	M	30	OR	1080	M	31	TM
1081	M	32	HN	1085	M	24	BN	1089	M	24	HN
1093	M	26	HN	1094	M	27	KN	1101	M	30	HN
1104	M	28	TL	1108	M	34	OR	1266	M	23	KN
1267	M	20	TL	1268	M	23	BN	1269	M	25	HN
1270	M	24	HN	1271	M	25	TL	1272	M	24	TL
1273	M	24	TL	1274	M	24	HN	1275	M	26	HN
1276	M	26	OR	1278	M	24	AS	1279	M	25	AS
1280	M	25	TL	1281	M	25	TL	1282	M	24	HN
1283	M	24	HN	1284	M	25	TL	1285	M	25	TL
1286	M	23	HN	1287	M	23	HN	1288	M	25	HN
1289	M	25	TL	1290	M	40	AS	1291	M	24	HN
1292	M	24	HN	1293	M	26	HN	1294	M	24	HN
1295	M	24	HN	1296	M	38	GJ	1297	M	24	HN
1298	M	31	OR	1299	M	33	OR	1300	M	38	OR
1301	F	26	HN	1302	M	22	ML	1303	M	30	OR
1304	M	24	TL	1305	M	25	TL	1306	M	28	OR
1307	M	29	OR	1308	M	24	MW	1309	M	24	MW
1310	M	25	OR	1311	M	25	OR	1312	M	25	BJ
1313	F	54	BJ	1314	M	25	TL	1315	M	28	TL
1317	M	25	BN	1318	M	24	TM	1319	M	24	TM
1320	M	33	HN	1321	F	30	HN	1322	M	25	HN
1323	M	24	HN	1324	M	25	TL	1325	F	23	TL
1326	M	26	HN	1327	M	42	HN	1328	M	24	AS
1329	M	22	ML	1330	M	23	ML	1331	M	23	ML
1332	M	23	KN	1333	M	20	ML	1334	M	20	ML
1335	M	22	ML	1336	M	23	ML	1337	M	24	HN
1338	M	22	HN	1339	M	30	ML	1340	M	24	ML
1341	M	36	ML	1342	M	27	ML	1343	M	23	HN
1344	M	25	HN	1345	M	23	HN	1346	M	23	HN
1347	M	23	ML	1348	M	23	ML	1349	M	28	ML
1350	M	27	ML	1351	M	23	ML	1352	M	23	ML
1355	M	26	ML	1358	M	25	TL	1361	M	28	TL
1362	M	33	AS	1365	M	36	OR	1368	M	29	HN
1372	M	27	MR	1375	M	26	HN	1376	M	50	OR
1377	M	31	HN	1380	M	25	HN	1390	M	32	OR
1391	M	32	TM	1392	M	28	HN	1393	M	25	BN
2014	F	25	TM	2034	F	23	ML	2040	F	26	KN
2041	F	40	KN	2046	F	26	HN	2047	F	27	HN
2083	F	28	TM	2100	F	29	AS	2102	F	32	AS
2103	F	27	HN	2130	F	45	BN	2131	F	18	TL
2132	F	20	HN	2133	F	55	KN	2134	F	26	KN
2135	F	40	KN	2136	F	26	TM	2137	F	22	AS
2138	F	23	TM	2139	F	23	ML	2140	F	23	ML
2141	F	23	ML	2142	F	23	ML	2143	F	40	ML
2144	F	36	ML	2145	F	35	ML	2149	F	23	BN
2148	F	24	HN	2150	F	30	AS	2152	F	25	HN
2154	F	31	HN	2161	F	25	BN				

- Part II: It contains 100 speaker pair database (two speaker recognition database). The conversation data of 100 speaker pair is stored along with the label files, which indicate the segments spoken by each speaker.

The complete database is stored in the following directory structure:

IITG-MV Phase III Speaker Recognition Database Part II (two speaker recognition)
→conversational style two speaker recognition→Session1(Session2)→Speech Files (along with the Label file).

5 IITG MV Phase-IV speaker recognition database

The fourth phase of data collection was done with the aim of facilitating the development of a GMM-UBM based online speaker verification system, to be ported on the telephone network. We used the Asterisk software and a telephony card to connect to the local PSTN (Public Switched Telephone Network). We were provided with a PRI line by BSNL for making and receiving calls on the network.

A 24 port PCI slot type telephone Card was used to connect to the local BSNL Telephone Exchange, through the PBX containing PRI card. A separate server was used, containing the telephone Card, to connect to the telephone lines. Asterisk software was installed on the server to use the card for building an IVR (Interactive Voice Response) System needed for data collection. Asterisk, as a tool, can be used for building a variety of telephone applications including Private Branch Exchanges (PBXs), Automatic Call Distribution Systems (ACDs), VoIP Gateways and other related communication projects. Here, we used Asterisk to build an IVR system for collecting data from distant subjects over the telephone network. The subjects were guided by the prerecorded IVR dialogues to give their speech data. The IVR System responded to their telephone keypad inputs, during the process of their data recording.

5.1 Speech data collection

5.1.1 Part-I

The whole data collection effort can be divided into two parts. In Part-I, the conference call facility of the mobile phone was used to record the speech data. The facilitator in the Laboratory first dialed the number of the BSNL line connected to the server. A welcome message stored in the server is played and asks to press a keypad digit to transfer the call to a particular extension number. After that, a PHP script file saved at the called extension number is executed. The script file is written to execute the desired call flow as shown in Fig. 7.

Before the call flow is executed, the script file first asks the facilitator to press the speaker ID of the subject whose recording is to be taken. This speaker ID is unique, and is chosen by the facilitator for each subject. Then, the facilitator calls the subject and adds him/her to the conference call. The script file then plays a series of instructions to the subject and successively stores his/her responses. The call flow during the entire recording is described by the flow diagram as shown in Fig. 7.

There is a total of 08 instructions played to the subject during the entire recording. Apart from these instructions, there is a total of 15 sentence files also that are played to the subject and have to be repeated by the subject. The instructions and sentence files are as follows:

- Please tell your full name.
- Please tell your father or husband's name.
- Please tell your mother or wife's name.
- Please tell your permanent address.

- Please tell your communication address, if it is same as permanent address, please repeat it.
- Please tell your mobile number.
- Please repeat each of the following sentences three times:
  - All cats are grey in the dark.
  - Come in the evening.
  - Give me a bucket of water.
  - Kanyakumari is a beautiful city.
  - Don't ask me to walk like that.
- Please read a text of your choice for three minutes. The text can be in any language and from any source, be it a newspaper, a magazine or any other book.

After all these instructions, a thank you message is played to the subject. A total of 23 wav files are played to the subject including eight instruction and fifteen sentence files. Subject responses to all the files are recorded in a subject specific folder, as shown in the flow diagram given in Fig. 7.

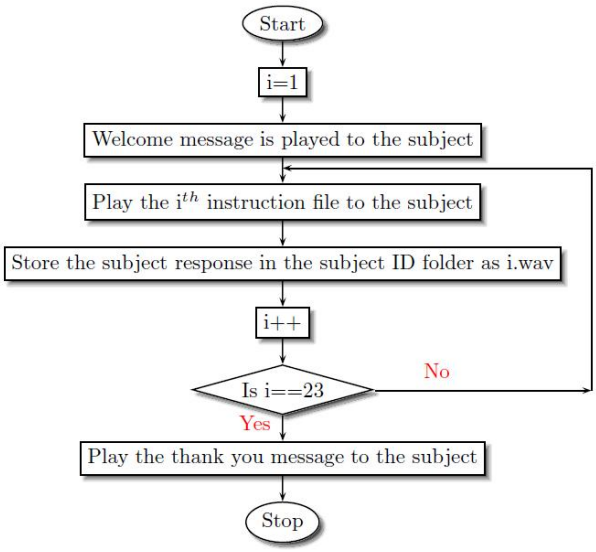

Fig. 7. Call flow diagram for recording Phase-IV of speaker recognition database

Using the above procedure, we collected data from 55 subjects covering all the major regions of the country. The collected speech files had to be used to aid in making the Universal Background Model (UBM). Thus, in order to get the desired variability, the protocol was to collect data of around 5 subjects from each region/state. The major regions covered in this way include Assam (North-east), Delhi/Punjab (North), Uttar Pradesh/Madhya Pradesh (representing Hindi speaking heartland) Karnataka, Tamil Nadu, Kerala, Andhra Pradesh (all representing south), Maharashtra and Gujarat (to represent west).

Since the data had to be used in making the UBM, only one session of recording was done for each subject. This whole data was combined with the Phase-III database to make a UBM. In previous Phase-III database out of 200 speakers 153 speakers were new and not common in any of the databases collected so far (Phase-I and Phase-II). We took these 153 speakers and combined with the 55 speakers collected in Phase-IV Part-I to make a 208 speaker UBM. In the 153 speakers data set there were only few females. So to ensure gender equality while collecting data for 55 speakers, we tried to increase the percentage of females. Our 55 speaker data set contains 37 females (67.28 %). The total data would be around 3.75 hours.

5.1.2 Part-II

Part-II of the data collection effort comprised of collecting data using a different IVR system. We implemented an IVR system that gives a unique speaker ID to each of the speaker and asked subjects to give 3 minutes of their reading style speech data for training the system. We collected data for 89 speakers this way.

5.1.3 Part-III

Part-III of the database consists of the the claim of the speakers against the speaker ID for testing. The speakers were asked to read a text for 30 seconds and this data is used for testing. Different trials of speakers were made against their speaker ID as genuine trials and against other speaker ID as imposter trials.

In total, we have collected around 7 hours of data, in Part-I,Part-II and Part-III of the entire Phase-IV of our data collection effort.

5.2 Processing of the data

In Phase-IV data collection (Part-I), after the entire recording is over, all the recorded responses (23 in number) are concatenated to form a single wav file. Thus, we have a specific wav file of around 4-5 minutes, for each subject, representing an entire single recording. All the files are stored on the server itself within their specific folders. Same way, the wav files collected during training the system, in Part-II of the data collection effort are stored on the server in their separate respective folders. The wav files recorded for testing the system,in Part-III are saved in two parts namely genuine and imposter trials. Further accepted and rejected files are stored separately under the two different trials.

5.3 Nomenclature for speech files

After processing, all of the speech files were named with a unique identification code which contains fields corresponding to speaker's ID , session, sensor, environment, language and style. The nomenclature of speech files of conversational style one speaker recognition database of Phase-IV follow the same as that of Phase-I and Phase-II. But the nomenclature of speech files of testing database of Phase IV is little different and as follows:

- The speech data files of testing has a serial claim number at the end of 13 digit nomenclature as mentioned.
- For details about different fields in the unique identification code, refer Table 3.

Example:9001_AM03MENR302.wav is the 302 numbered claim in the system made by a speaker against speaker id 9001, first session, mobile phone, multi-environment and English language in reading mode is used for recording.

5.4 Organization of the database

After giving appropriate names,each of the recorded speech files were stored in structured manner. The phase IV database is divided in three parts, which are:

- Part I: The part I of of phase IV database is the database collected for the purpose of creating the Universal Background Model (UBM). Speech files under this section is of total 4-5 minutes of duration for each speaker.
- Part II: The part II of phase IV database is collected for training the system. Speech files under this section is of 3 minutes of duration.

Table 11
The list of the languages and their occurrences in the IITG MV Phase-IV dataset under subjects' favorite language category

Language	Occurrence	
	Phase-IV Part I	**Phase-IV Part II**
Telugu	2	-
Oriya	4	-
Bengali	2	-
Assamese	1	-
Gujarati	2	-
Kannada	1	-
English	43	60

- Part III: The part III of phase IV database is collected for testing purpose. The data here is again subdivided into two different types of trials,namely Genuine and Imposter. And both these divisions have two more subdivisions namely, accepted and rejected. Speech files under this section is of 30 seconds of duration.

The complete database is stored in the following directory structure:

IITG-MV Phase IV Speaker Recognition Database→Part I (Part II, Part III)→UBM (Train, Test)→ Speech Files.

Data under Test directory is further organized as: Genuine(Imposter)→Accepted(Rejected)

6 Study of different mismatches on speaker verification system performance using IITG-MV database

The speaker verification (SV) research in recent times is concentrated towards addressing the degradation in the performance due to the mismatch between training and testing conditions. In this section we report the results of the initial studies conducted using the newly collected IITG-MV database to quantify the effect of the mismatches in sensor, language, style of speech and environment between the training and testing phases.

6.1 Experimental setup

For the experiments, a speaker verification system was developed which uses Gaussian mixture models (GMMs) created by *maximum a-posteriori* (MAP) adaptation of means of a universal background model (UBM) using the feature vectors of the training data for modeling the speakers [Reynolds et al., 2000]. The UBM was built with a mixture of 1024 Gaussian components with diagonal covariance matrices. The verification score is calculated by finding the average log likelihood of the test data features with the claimed speakers model. 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 [Ganchev et al., 2005]. 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 evaluating the speaker verification system performance on the IITG MV database, a set up similar to that of the NIST 2003 SRE evaluations [Martin, 2003] is made. In this setup, speech utterances of 2 min duration are used for training the speaker models and utterances of 30-45 s are used for testing. There are 10 test segments for each speaker, each of which is tested against 10 false claims and one true claim. The equal error rate (EER) noted from the detection error trade-off (DET) curve [Martin et al., 1997] is used to quantify the speaker verification performance. All the experiments were done using the HTK software [Young et al., 2006]

6.2 *Study of sensor mismatch*

Sensor mismatch is one of the main challenges in the area of speaker recognition and verification [Reynolds, 1996].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 (DVR). To study the effect of these sensors on the performance of the system, we kept the other variabilities like language, style of speech and environment of recording fixed across the training and testing data. For this experiment we have used following conditions: the environment is office, the language is English and the style of speech is conversation. The training data contains 100 speech utterances of 2 minutes each, derived from the conversations. These speech segments have been used to model the corresponding speakers by MAP adaptation of a gender independent UBM. The UBM was made using 6 hours of speech data from 50 different speakers from Phase-II not used in the speaker verification study. The test data set contains 1000 segments of speech of length varying from 30 to 45 seconds, 10 from each speaker. Each test segment is tested against 11 models out of which one may be a true trial. This makes a total of 11,000 test trials with a true trial to false trial ratio approximately equal to 0.1. All the possible training-testing combinations were tried with the data from five different sensors.

Table 12
EERs for sensor matching and mismatch cases, environment: office, language: English, style of speech: conversation.

Sensor	ERR (in %)				
Train ↓ Test →	Headset	Mobile phone-1	Mobile phone-2	Tablet PC	DVR
Headset	3.82	11 .01	13.82	14.01	18.7
Mobile phone-1	9.54	9.03	9.17	9.83	20.74
Mobile phone-2	9.52	9.12	9.01	9.79	21.35
Tablet PC	10.12	13.18	12.94	9.15	23.91
DVR	18.65	21.02	21.28	18.94	14.48

The performances in terms of EERs are shown in Table 12. The best EER of 3.82 % is observed for the matched case with headset data which is the cleanest data among that of the considered sensors. In the other sensors matching cases, the DVR shows the highest EER of 14.48 % and the mobile phones and Tablet PC mic show almost same EER which falls in between that of the other two and is around 9 %. For the sensor mismatch cases, that is, when data from different sensors are used for training and testing, it can be observed that, considering any two sensors, the matching case always gives better performance than the mismatch cases irrespective of the quality of the speech data. This leads to the conclusion that the mismatch in training and testing sensors results in large degradation in performance.

6.3 *Study of language mismatch*

In the IITG-MV Phase-I dataset each of the 100 speakers speaks in two languages: English and the subject's favorite language. To study the effect of language in speaker verification performance, we have used the data from 100 speakers recorded in the office environment. Training was done using English language and testing using subject's favorite language/mother tongue and vice versa. The other settings, sensor used for recording and style of speech are kept fixed to headset microphone and conversation mode, respectively, during the training and testing phases. The configuration of the SV system remains same as that used for the previous experiments. 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 EERs in Table 13. 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. Where as, in case of mismatched language cases, the performances degrade by about 1.2 % in EER with respect to their matched case performances.

Table 13
EERs for language matching and mismatch cases, environment: office, style of speech: conversation, sensor: headset microphone.

Language	ERR (in %)	
Train ↓ Test →	English	Favorite
English	3.82	5 .01
Favourite	5.34	3.91

6.4 Study of 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 EERs in Table 14. It is noted that, for mismatched testing cases a performance degradation of about 4% in EER compared to that of the matched case.

Table 14
EER for for style matching and mismatch cases, environment: office, language: English, sensor: head set microphone. Abbreviations: Conv-Conversational

Style of speech	ERR (in %)	
Train ↓ Test →	Read	Conv
Read	2.87	6 .71
Conv	8.54	3.91

6.5 Study of 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, were used. As across Phase-I and Phase-II there were 50 common speakers, the study was done on the set of those 50 speakers rather than on 100 speakers set used for the previous 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 uncontrolled environments data are given in terms of EERs in Table 15. It is to note that the change in recording environment affects the performance of the speakers verification system largely. The performance degradation due to mismatch in environment is large when the models are trained using a clean data and test using noisy data.

Table 15
EER for for environment matching and mismatch cases, language: English, style: conversation, sensor: head set microphone. Abbreviations: Multi- multi-environments

Environment	ERR (in %)	
Train ↓ Test →	Office	Multi
Office	3.92	10 .43
Multi	11.01	8.98

6.6 Speaker verification system performance on wireless mobile channel data

One of the major areas of application for speaker verification is person authentication though mobile phones. The wireless channel can cause degradation in the quality of data due to packet lose. It is important to study the effect of mobile wireless channel on the performance of SV system [Reynolds et al., 1995]. Here we use the Phase-II dataset of the IITG MV database for exploring the impacts of wireless channel on speaker verification performance. In Phase-II, for collecting speech over mobile wireless channel, during the recording, 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. In addition to the mobile channel data, this dataset contains speech recorded in parallel using the headset microphone, the tablet PC microphone, the digital voice recorder and an offline mobile phone (mobile phone used as a voice recorder). For the study of effect of wireless channel, we have used the data collected using the headset, offline mobile phone and the online mobile phone. The other conditions of the data were kept as English language and conversational style. The evaluation protocol followed was the NIST 2003 SRE evaluation plan for primary task and the other system parameters were kept same as that of the baseline system. Models for 100 speakers were created using the data collected from the headset microphone data. The testing was done using the second session data collected using headset microphone, the offline mobile phone and the online mobile phone. In addition to these, we have also conducted experiments to benchmark the Phase-III dataset of the IITG MV database using the same SV system setup. The Table 16 shows the results of the various experiments conducted. As explained earlier, the data for Pahse-IV dataset was collected using an online SV system through telephone channel. The EER performance of the online SV system was observed as 21.3 %.

Table 16
EER for for offline and online recording cases, language: English, style: conversation, sensor: mobile phone.

Condition	EER in %
H01 Train H01 Test	8.98
H01 Train Offline mobile(M04) test	13 .43
H01 Train Online mobile(M03) test	15 .21
Phase-III data, Part-II: Mobile Train- Mobile Test	11 .17

7 Summary

A recently collected speech database for the purpose of developing a robust speaker recognition system in Indian context is reported. The main objective was to collect a database that contains most of the variabilities that occur in Indian scenario. The 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.

Our study exploring the impact of mismatch in training and test conditions with collected data infers that for the mismatch in sensor, speaking style, and environment, there is significant degradation in performance compared to matched case. However, the effect of mismatch in language is found to be relatively less. 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. The authors sincerely thank the efforts of Mr. Akhilesh Shukla and Mr. Sumit Shukla for their effort towards the collection and processing of database.

References

Campbell, J. P., Reynolds, D. A., 1999. Corpora for the evaluation of speaker recognition systems. In: Proceedings of International conference on acoustics, speech and signal processing 1999, (ICASSP '99).

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., May 2011. Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing 19 (4), 788 –798.

Doddington, G., Nov. 1985. Speaker recognition-identifying people by their voices. Proc. IEEE 73 (11), 1651–1664.

Ganchev, T., Fakotakis, N., Kokkinakis, G., 2005. Comparative evaluation of various mfcc implementations on the speaker verification task. In: Proc. SPECOM. pp. 191–194.

Haris B C, Pradhan, G., Misra, A., Shukla, S., Sinha, R., Prasanna, S. R. M., 2011. Multi-variability speech database for robust speaker recognition. In: Proceedings of National Conference on Communications. pp. 1–5.

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

Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., 1997. The DET Curve in Assessment of Detection Task Performance. In: Proceedings of Eurospeech '97. Rhodes, Greece, pp. 1895–1898.

Patil, H., Basu, T., 2008. Development of speech corpora for speaker recognition research and evaluation in indian languages. International Journal of Speech Technology 11, 17–32.

Patil, H., Prakash, D., Kar, B., Bhatta, B., Basu, T., dec. 2006. Corpora for speaker recognition research and evaluation in oriya. In: Proceedings of IEEE International Conference on Industrial Technology. pp. 2217 –2222.

Reynolds, D., may 1996. The effects of handset variability on speaker recognition performance: experiments on the switchboard corpus. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing 1996 (ICASSP '96). Vol. 1. pp. 113 –116 vol. 1.

Reynolds, D., Zissman, M., Quatieri, T., O'Leary, G., Carlson, B., may 1995. The effects of telephone transmission degradations on speaker recognition performance. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing 1995 (ICASSP '95). Vol. 1. pp. 329 –332 vol.1.

Reynolds, D. A., may 2002. An overview of automatic speaker recognition technology. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing 2002 (ICASSP '02). Vol. 4. pp. IV–4072 –IV–4075.

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

www.speech.kth.se/wavesurfer/, -. Wavesurfer.

Yin, S.-C., Rose, R., Kenny, P., 2007. A joint factor analysis approach to progressive model adaptation in text-independent speaker verification. IEEE Transactions on Audio, Speech, and Language Processing 15 (7), 1999–2010.

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