International Journal of Research in Advanced Engineering and Technology
ISSN: 2455-0876; Impact Factor: RJIF 5.44
www.engineeringresearchjournal.com
Volume 2; Issue 3; May 2016; Page No. 162-164

Protocol for database collection in Indian scenario

Pardeep Sangwan

Maharaja Surajmal Institute of Technology, GGSIPU, New Delhi, India

Abstract

There are various challenges in speaker recognition mainly due to lack of speaker databases having speech data addressing the several issues affecting the efficiency of the speaker recognition systems like session variability, channel variability, with-in and between speaker variabilities. Due to scarcity of the speaker databases, it is intended to collect a database which can address these problems of speaker recognition systems. This paper describes a protocol developed for collecting database of speech samples in Indian scenario for further research work. First, the requirements and the rationale for these requirements are described. Then, description of the technical equipment and setup employed is given. Finally, the post-processing of the recordings is explained to prepare them for inclusion in the database.

Keywords: Speaker database, Channel mismatch, Speaking Styles

1. Introduction

Speaker recognition is a technology which is used to authenticate persons from their speech samples. One of the most important challenges in speaker recognition stems from inconsistencies in the different types of speech samples and their quality. One such problem, which has been the prime focus of researchers, is the problem of channel mismatch, in which the speech data has been collected using one apparatus and the test sample has been recorded by a different channel. It is important to note that the sources of mismatch vary and are generally quite complicated. There could be any combination and usually are not limited to mismatch in the handset or recording apparatus, the network capacity and quality, noise conditions, illness related conditions, stress related conditions, transition between different media, etc. Some approaches involve normalization of some kind to either transform the data (raw or in the feature space) or to transform the model parameters. Beigi [1] discusses many different channel compensation techniques in order to resolve this issue. Vogt, *et al*. [2] provided a good coverage of methods to handle modelling mismatch. One such problem is to obtain ample coverage for the different types of phonation in the training and enrolment phases, in order to have a better performance for situations when different phonation types are uttered. An example is the handling of whispered phonation which is, in general, very hard to collect and is not available under natural speech scenarios. Phonation deals with the acoustic energy generated by the vocal folds at the larynx. The different kinds of phonation are unvoiced, voiced, and whisper. Unvoiced phonation may be either in the form of nil phonation which corresponds to zero energy or breathe phonation which is based on relaxed vocal folds passing a turbulent air stream. Majority of voiced sounds are generated through normal voiced phonation which happens when the vocal folds are vibrating at a periodic rate and generate certain resonance in the upper chamber of the vocal tract. Another category of voiced phonation is called laryngealization (creaky voice). It is when the arytenoid cartilages fix the posterior portion of the vocal folds, only allowing the anterior part of the vocal folds to vibrate. Yet another type voiced phonation is a falsetto which is basically the un-natural creation of a high pitched voice by tightening the basic shape of the vocal folds to achieve a false high pitch. In another view, the emotional condition of the speaker may affect his/her phonation. For example, speech under stress may manifest different phonetic qualities than that of, so-called, neutral speech [3]. Whispered speech also changes the general condition of phonation. It is thought that this does not affect unvoiced consonants as much.

Also, the phonation undergoes certain changes when the speaker is under stressful conditions. Bou-Ghazale, *et al*. [3] have shown that this may affect the significance of certain frequency bands, making MFCC features miss certain nuances in the speech of the individual under stress. They propose a new frequency scale which it calls the exponential-logarithmic (expo-log) scale. Although research has generally shown that cepstral coefficients derived from FFT are more robust for the handling of neutral speech [4], Bou-Ghazale, *et al*. [3] suggest that for speech, recorded under stressful conditions, cepstral coefficients derived from the linear predictive model [1] perform better.

In order to objectively address these challenges one must use a database containing samples from a large number of speakers that constitute a representative sample of the relevant population. If there is variability at the source or in the extraction of information that constitutes a sample, then in order to calculate the degree of validity and reliability of the system, the database must contain multiple samples from each object [5].

Most of the speaker recognition databases available are in American English and do not cover Indian languages and environmental conditions. To overcome this constraint, the creation of multi-device, multi-lingual and multi-environment speech database for speaker recognition tasks is described here.

This paper presents a protocol developed for collecting database of speech samples in Indian scenario for further research work. First, the requirements and the rationale for these requirements are described. Then, description of the technical equipment and setup employed is given. Finally,

the post-processing of the recordings is explained to prepare them for inclusion in the database.

2. Requirements

In designing the protocol, there were two basic requirements. In this section each requirement and its rationale is explained.

a) Requirement 1

The first requirement is that the database contains recordings of each speaker using different speaking styles, that these be typical of speaking styles found in speech samples, and that they be elicited as natural speech without requiring the speakers to role play. The speaking styles of recordings often differ; for example, one recording may be of a lively telephone conversation and another of relatively subdued responses to police interview questions. It would be impossible to cover all possible gradations of speaking styles, and eliciting genuine examples of different emotional states would be very difficult. Therefore, three speaking styles that we believe are common and which are therefore likely to be useful in a large proportion of cases. These styles are: (1) informal conversation; (2) information exchange task over the telephone (including exchange and confirmation of numbers and letters); (3) reading of written material. These tasks are described in detail below. Each task lasts approximately ten minutes. We expect to obtain four or more minutes of speech from each speaker on each task. The tasks were designed such that the speakers are put into particular situations and, without any special instructions, simply have to act normally and complete the tasks as they themselves would naturally do them.

b) Requirement 2

The second requirement is that the database be usable for research work involving recording- and transmission-channel mismatch. Channel mismatch is common in speaker recognition tasks. Channel mismatch introduces differences between recordings that are confounded with within-speaker and between-speaker differences. One solution is to collect high-quality speech database through different channels allowing for fine-grained examination of the effect of each channel.

3. Database collection setup and Equipment used

In this section the equipment and its setup is described. The details of the particular equipment employed are included because many people ask for these details, and in most cases substantial time is invested in researching which equipment would best suit the requirements. This is not to preclude that other equipment could be substituted.

The database for speaker recognition is collected from five different sensors placed parallel to one another to obtain a wide variety of channel variability. The database is collected in English, Hindi and mother tongue of the subject as detailed in table1. Recording is done in three speaking styles and in uncontrolled environment such as laboratories from 40 speakers out of which 30 are male and 10 are female speakers. The details of different speaking styles are given in table 2. The larger the number of recordings per speaker, the better will be the ability to estimate the reliability of the speaker recognition system. The speech data samples are recorded in a laboratory with a fan and an air conditioner powered on and in the presence of students. The data was collected in parallel with a headset microphone connected to a PC, the built-in microphone of the same PC, one mobile phone with voice recording facility, one collar mic and one Bluetooth enabled headphone-mic. The technical details of the sensors are given in table 3. The speech data was contributed by 30 male and 10 female volunteers among students, staff and faculty at MSIT.

Table 1: Languages of speech samples

S. No.	Language	ID	No of Speakers	
			Male	Female
1	English	01	30	10
2	Hindi	02	27	10
3	Tamil	03	6	3
4	Punjabi	04	9	4
5	Garhwali	05	7	1
6	Kumayuni	06	4	0
7	Rajsthani	07	4	2

Table 2: Different Speaking Styles Used

S. No.	Speaking Style	ID
1	Reading	RD
2	Conversational	CN
3	Information Exchange	IE

Table 3: Details of Sensors used

S. No.	Sensor	ID
1	Laptop Mic (Lenovo Z516)	01
2	Headphone Mic (Philips)	02
3	Bluetooth Headphone Mic (Sony BH-503)	03
4	Mobile1 (Micromax A350)	04
5	Collar Mic (Studio Master)	05

The microphones are connected via high-quality cable to a USB Audio Capture card, which is in turn connected to a computer. Software such as Cool Edit Pro and Audacity is used to record the incoming speech signal. It is important that the software provide a live display of the incoming signals so that the researcher can adjust the recording levels. Usually the microphone gains are set at the beginning of the recording session and not changed thereafter. Peak limiters are switched off to avoid the distortion that they introduce to the signal, so it is important to avoid clipping by not setting the gain too high. For a good-quality recording taking advantage of the number of bits available for amplitude encoding, it is also important that the gain not be set too low. High-quality headphones such as Sony BH-503 or Bh-505 are used as Reference Headphones so that they can monitor the recordings for extraneous noises and other problems due to factors such as poorly placed microphones, speakers playing with the microphone cable, etc. and take corrective action as needed. The recording equipment is set up and recording is started at the beginning of the recording session and not switched off until all three tasks have been completed (an alternative would be to stop recording and save after each task, there are pros and cons to each approach). It is advisable to use a relatively fast/powerful computer and run some pilot trials to make sure that it does not drop any of the incoming signals (some older laptops have been reported to have buffering problems and to drop frames from the incoming signal). It is advisable that the

computer have a battery backup so that data are not lost due to a power cut. We make the recordings at 44.1 kHz frequency sampling and 16 bit amplitude sampling, and save them as raw PCM wave files. File naming conventions are described below under post-processing of recordings. It may be advisable to have an external storage device connected to the computer to which a backup of each recording is automatically immediately made and a second external storage device to which a second backup is manually made at the end of each day.

The nomenclature adopted for the recorded files is given in Table 4.

The first three alpha-numeric characters are the unique speaker ids allotted to individual speakers. Out of these three alphabets, first alphabet is for the gender of the speaker and other two are number allotted to individual speaker. Next two characters are for channel/sensor used for that particular recording. Then, next two characters are describing the speaking style id and the last two depicts the language of utterances recorded.

Table 4: Nomenclature of recordings

Alpha-numeric Characters								
1	2	3	4	5	6	7	8	9
Speaker ID			Channel ID		Style ID		Lang. ID	

4. Conclusion and Future scope

It was tried to incorporate several issues related to the speaker databases like channel variability, speaking style variability and variability due to different languages of speech samples which can greatly affect the efficiency of the speaker recognition task. The database collected is having all these diversities but to a smaller extent. In the next phase of data collection these variabilities will be enhanced and a large number of speakers will be included to get a robust speaker database.

It is believed that the description of this database collection protocol is of value to other researchers. Some researchers may wish to adopt the protocol essentially as is. For others, the description of our protocol may help advance their thinking about their own database requirements and practical database collection issues.

5. References

1. Homayoon Beigi. "Fundamentals of Speaker Recognition", *Springer*, New York, 2011. ISBN: 978-0-387-77591-3.
2. Robbie Vogt, Sridha Sridharan. "Explicit Modelling of Session Variability for Speaker Verification", *Computer Speech and Language*. 2008; 22(1):17-38.
3. Sahar E Bou-Ghazale, John HL Hansen. "A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech under Stress", *IEEE Transactions on Speech and Audio Processing*. 2002; 8(4):429-442.
4. Dhanalakshmi P, Palanivel S, Ramalingam V. "Classification of Audio Signals using a ANN and GMM", *Applied Soft Computing*. 2011; 11(1):716-723.
5. Morrison GS. "Measuring the Validity and Reliability of Forensic Likelihood-Ratio Systems", *Sci. & Justice*. 2011; 51:91-98. doi:10.1016/j.scijus.2011.03.002.