Global Journal of Enterprise Information System DOI: 10.18311/gjeis/2017/16123

A Structured Approach towards Robust Database Collection for Speaker Recognition

Pardeep Sangwan[1*] and Saurabh Bhardwaj[2]

[1]*Department of ECE, Maharaja Surajmal Institute of Technology, New Delhi, Delhi, India; sangwanpardeep@gmail.com*

[2]*Department of EIE, Thapar University, Patiala, Punjab, India; bsaurabh2078@gmail.com*

Abstract

Speaker recognition systems are classified according to their database, feature extraction techniques and classification methods. It is analyzed that there is a much need to work upon all the dimensions of forensic speaker recognition systems from the very beginning phase of database collection to recognition phase. The present work provides a structured approach towards developing a robust speech database collection for efficient speaker recognition system. The database required for both systems is entirely different. The databases for biometric systems are readily available while databases for forensic speaker recognition system are scarce. The paper also presents several databases available for speaker recognition systems.

Keywords: Age Variability, Recording Instruments, Session Variability, Spoofing, Whispering

Paper Code: 16123; **Originality Test Ratio:** 7%; **Submission Online:** 20-May-2017; **Manuscript Accepted:** 24-May-2017; **Originality Check:** 30-May-2017; **Peer Reviewers Comment:** 01-June-2017; **Double Blind Reviewers Comment:** 11-June-2017; **Author Revert:** 16-June-2017; **Camera-Ready-Copy:** 18-July-2017

1. Introduction

Speech is a means not only to convey message through words spoken but speech also has information about the speaker which can be used for identifying the speaker. On the basis of application speaker recognition has two parts namely: speaker identification and speaker verification. The combination of these two systems must be able to identify a speaker from the group of known speakers along with classifying a large number of unknown speakers to an invalid speaker category. Speaker recognition used in the field of Forensic Science is called Forensic speaker Recognition (FSR). It is a system to decide whether two speech samples are uttered by same person or not. The main difference between FSR and speaker recognition is the database on the basis of which recognition is to be done. In former database constitutes primarily the speech samples of non-cooperative speaker recorded in a noisy environment (i.e. samples of poor quality) while latter uses the speech samples from cooperative speaker recorded generally in a clean environment. This difference drastically changes the requirements of the speaker recognition system. As discussed by Haris et al.[1], the standard speech database has an important role in developing and evaluating a speaker recognition system. The availability of speech databases may be one factor responsible for the progress made in the field of speaker recognition.

2. Important Factors for Speech Database Collection

To collect a speech database for biometric speaker recognition the main factor to be taken into consideration is session variability as speakers are generally cooperative in this type of speaker recognition but in case of FSR along with session variability several other factors are ought to be considered because of the fact that uncooperative speakers are encountered almost every time in FSR. Some of the important factors are shown in Figure 1.

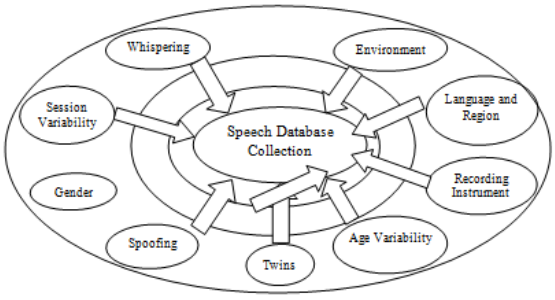


Figure 1. Important Factors for a Robust Speech Database collection for both systems.

*Author for correspondence

There are many important factors to be considered while collecting database for both biometric and forensic speaker recognition. Some of these factors are discussed below:

Session Variability: Session variability also known as Inter-session variability refers to all the phenomena causing variations in two recordings of same speaker.[2] In other words, two speech samples recorded by the same person are mismatched and could not be recognized by the system[3]. There are several factors responsible for inter-session variability such as:

- ***Transmission channel*:** This is one of the major factors responsible for inter-session variability. This happens due to the difference in the characteristics of the transmission channels. For example, the speech signals recorded from the voice transmitted over mobile phone i.e. wireless channel may have different characteristics as compared to the speech samples from transmission over landline phone i.e. wired channel in spite of the fact that both the utterances are from the same speaker.[3]
- ***Transducer characteristics*:** Transducer is a device which converts one type of signal to another type of signal. In case of speech, audio signal is converted into electrical signal by the transducer known as microphone. This electrical signal is then processed and stored as speech sample. Various types of microphones are available in the market like carbon, electret, laser, fiber optic, ribbon, MEMS, and liquid microphone and many more. Quality of speech sample also depends on type of microphone used. When training and testing devices have different kind of microphones then the speaker recognition system may not recognize the target speaker accurately.
- ***Environmental noise*:** This is another factor which can cause inter-session variability as amount of noise may be different in different sessions of recording. Noise may be due to several reasons like traffic, people talking in the near vicinity of the recording place, some industry nearby or many more.[4]
- ***Intra*-speaker *variability*:** Speech samples from the same speaker may also cause inter-session variability because the acoustic features of the speech sample largely depend on the age, physiological and psychological health, emotional state etc. of the speaker. Thus, if there is a large time span between training and testing sample recording there may be huge change in the features of the voice of the same speaker. Furthermore, if target speaker is ill or in a different emotional state, then also features may change and cause inter-session variability.[5]

For robust speaker recognition, inter-session variability must be compensated. Several techniques have been proposed for compensation like feature mapping, feature warping, H-norm, T-norm and many more.

Gender: Biometric systems use the traits like fingerprint, iris, palm-print, face, hand-geometry and voice etc. to recognize individuals[6]. Along with these traits, some ancillary information of the user can also be used to design a reliable biometric system which is also user-friendly. This ancillary information could be height, gender, age, eye color or weight of the individual person. These are known as 'Soft Biometric' traits. These traits can be continuous or discrete. Gender is discrete soft biometric trait. Though, use of soft biometric traits suffers from lack of distinctiveness still gender and age can be used to filter out a large biometric database. It is quite obvious that voice of male speakers and that of female speakers have great variations in context of several features of speech signal and hence can filter out large amount of undesirable data and save a lot of precious time.

Environment: The mismatching between training and testing conditions has become the prior concern of the researchers in the recent past. Various techniques have been developed to address this problem like speaker model synthesis, several normalization techniques; factor analysis, feature mapping and nuisance attribute projection. Many out of these techniques require parallel condition data, which is not contained by most of the publicly available speaker recognition databases.[7] For the robust speaker recognition the parallel condition data must be available as the surrounding conditions also affect the quality of the speech signal. For example, quality of speech sample recorded in a sound proof room will be much better than the speech sample recorded in a classroom or library. Sometimes noise in the speech samples is also desirable factor for designing a robust speaker recognition system, so that, the system can recognize the speaker even in the extreme conditions.[8] For this purpose a database should have the speech samples recorded in the different environments like a sound proof room, noisy class room, library, auditorium and market etc.

Language and Region: Speaker verification can be categorized as text dependent and text independent. In text independent system speaker is not constrained about what to speak for verification. It is an implicit verification where the speaker is performing some other task like registering a complaint or talking with customer-care executive. Most of the speaker recognition system works in a single-language environment. India is a highly multilingual country having large no of regional languages. Different languages have different syntactic and lexical rules governing the structure and pronunciation of that language. Furthermore, different regions have their own effect on the language and its acoustic features. Even the same language is spoken in different accent across the different regions of India. For example, English spoken in south India have different acoustic feature like pitch, energy, phenomes etc. as compared to the English spoken in north India. Hence, it is very important to analyze the effect of multiple languages as well as the effect of single language spoken

in different regions on speaker recognition system. So, a good database should have speakers of the same language from different regions as well as speakers of different languages. Recent advancement in databases of Indian languages are either limited to 4-5 languages or concentrated on a specific geographical area. The IITKGP-Multilingual Indian Language Speech Corpus (IITKGP-MLILSC) comprises of 27 Indian regional languages including 16 languages which are most widely spoken across India.

Recording Instrument: Recording of the speech samples can be done through various devices. Quality of the speech samples also depends upon the device used for recording. Some of the devices which can be used for a quality database are digital voice recorder, laptop, mobile phone, microphone and long distance call over phone.

Age Variability: The problem of ageing and variation of speech quality go hand in hand. The ageing effect increases with time and also the quality of the speech is more likely to be degraded with time. Speaker verification with these two parameters of ageing and variation in the speech sample quality is a very important problem. In any biometric system, the accuracy of the system degrades due to ageing effect. However, in speaker recognition, the effect of age variability has received marginal research attention. The acoustic changes to the voice due to physiological changes in the vocal mechanism have been extensively studied in. Most of the noticeable changes in voice take place in childhood and the old age but voice keep on changing progressively throughout adulthood. Out of these, the most important changes are (i) a downward shift in fundamental frequency, and (ii) a change in timbre. Generally in the speakers of age more than 60 years, the typical changes are instability of pitch and intensity of voice, and slowed rate of delivery. In speaker recognition system, this problem of ageing can be overcome by updating the database at regular intervals of time but for large-scale system, this solution degrades the security of the system and hence not a feasible solution at all. A better but complex solution is to adjust ageing-related changes automatically. The main problem in designing such a system is the lack of the database. The longitudinal speaker database which is covering a time span of more than three years is unavailable publically. A stacked classifier framework for improving performance of the long term speaker verification system via an ageing-dependent decision boundary using Trinity college Dublin Speaker Ageing (TCDSA) database is proposed in[9]. In TCDSA database, the main variability was ageing but variation in speech quality was also unavoidable over such a long time span. Thus, for a long term and large scale system, the database must have data of a speaker at different times or database should be updated in this manner regularly.

Spoofing: Text-dependent speaker verification systems have high accuracy due to short utterances but such systems are not much reliable against spoofing attacks. Basically there are four types of spoofing attacks: (i) impersonation, (ii) speech synthesis, (iii) voice conversion, and (iv) replay. Impersonation means a person tries to mimic someone else who is actually a genuine speaker. The effect of impersonation attack on speaker verification system and vulnerability of the system had been studied in[10]. In speech synthesis, the voice of the genuine speaker is synthesized using a speech synthesizer for spoofing the verification system. Voice conversion is an approach in which the voice of the attacker is converted automatically using a conversion function to mimic the voice of genuine speaker for spoofing the system. Replay is the most easily implementable, low technology spoofing attack approach, which do not need any speech processing techniques. In this the pre-recorded speech samples of the target genuine speaker are replayed using a playing device that can be a mobile phone, music player or any other player. Hence to avoid spoofing attacks, measures should be taken right from the first phase that is, the speech database collection phase.

Whispering: A lot of research has been done in the field of speaker recognition but researchers have focus on the recognition of 'cooperative speakers' only. Cooperative speakers are the speakers who want to be identified by the automatic speaker recognition system. But speakers, who don't want to get identified i.e. un-cooperative speakers, may lower their voice or try to change their speaking behavior to avoid recognition and fool the system intentionally. The effect of intentional speaking behavior modification on the speaker recognition is investigated by Kajarekar et al. and presented vulnerability in speaker recognition systems. The performance of the speaker recognition system degrades greatly if mismatch appears between training and test set. For example, the voice of a person varies when psychological or physiological conditions of the speaker changes. A disguised speech, produced psychologically and/or physiologically, known as whispered speech has been an area of interest for researchers in the recent past.[11] Along with the research on its acoustic features like formant frequencies, corresponding bandwidth and endpoint detection etc., its applications like its reconstruction, speaker recognition, and many more, have also been attracted the attention of the researchers. The main features of the whispered speech are: i) Exhalation is the source of excitation as vibration of the vocal cords are absent. ii) Whispered speech has very low SNR and hence, the surrounding easily effect on it. iii) The psychology of the enunciator is susceptible while whispering.[11]All the above aspects can affect the accuracy of the speaker recognition system greatly and thus, it is very important to have database of whispered speech for the development of an efficient speaker recognition system.

Twins: According to the statistics in, birth rate of twins is increasing with an average 3% per year since 1990. The number of identical twins is only 0.2% of the world's population but still it is

equal to the population of countries like Greece or Portugal. Due to this, there is an urgent need for a biometric system capable of accurately distinguishing between identical twins, who share same genetic code. Basically, there are two types of twins namely:

- Monozygotic (MZ) twins occur when one zygote is formed by fertilization of single egg, which divides into two separate embryos. These are the identical twins and shares 100% of their genes.
- Dizygotic (DZ) twins occur when two separate eggs are fertilized by two different sperm cells. These fraternal twins shares 50% of their genetic information.

Many studies are carried out on twin pairs with several research objectives. Some of them can be given as: (i) try to distinguish a speaker from his co-twin, (ii) try to find genetic component in the variation of certain acoustic features. The results have been shown that the twin pairs have different degree of similarity and dissimilarity in their voice and speech parameters. The results also depend on a particular twin pair under consideration as well as on the acoustic parameters selected for discrimination. Hence, several parameters have been considered for research to assess twin's (dis)similarities. But all the efforts are worthless if a database of twin pairs is not available and hence for a robust speaker recognition system as well as for forensic applications twins' database is very important.

Duration Variability: The length of the available unknown speech sample may be of very short duration as compared to the recorded speech samples in the database. This could adversely affect the recognition efficiency. This is shown by researchers through several experiments that increasing the duration of training data can improve the efficiency of the recognition system.

Different Speaking style and Situational mismatch: Unknown speech samples are recorded from different sources generally unknown to the speaker uttering them. Thus the speaking style could be same or entirely different than the stored speech database. The style can be normal conversation between two persons on telephone or it may be simply something read by the speaker or it may be an utterance in which speaker is yelling over someone. It quite possible that recorded speech is containing voices of more than two persons (meeting style) or may be someone disguising intentionally to create confusion. It may also be the part of police-suspect interview or may be some kind of information exchange over phone.

Different Stress level and Mental state of the speaker: Speech samples could result in error if the speaker is in a different mental state at time of two utterances to be compared. For example, if one sample is recorded when mental state of the speaker was normal and another speech sample is taken when speaker is mentally ill or under the impression of some sedative drug. Both samples can have large intra-speaker variability. The same could be caused by different stress level too irrespective of the fact that the stress could be physical mental or emotional.

Sparse background data: The background database required for the development of forensic speaker recognition model is sparsely available due to the legal problems in forensic database collection. Due to this reason, very less amount of forensic data is available to researchers till date and to the best of our knowledge none is available in the Indian scenario.

3. Available Databases

It is very difficult to overcome all the above mentioned problems in a single database. Many researchers have tried their best to collect database which can have samples as near to realistic forensic data as possible but could not succeed to address all the above said problems. Some of the very well-known databases around the world are presented in the table 1.

Table 1. Available Database for Forensic Speaker Recognition

S. No.	Name of Database	Language Used	Year of Release	Duration/ Size	No of Speakers		Type of Data/ Text used	EER	Speaking Styles used
					M	F			
1	FABIOLE[12]	French	2016	3100 utterance	130		News, Debate, TV Shows	2.5	Reading
2	NFI-FRITS[13]	Dutch	2014	4188 conversations	604		Conversational	12.1	Actual conversations intercepted by Police
3	CIVIL-CORPUS[14]	Spanish	2013	20 Hrs	28	32	Sentences/Digits/Word	-	Disguise/ Conversation/ Reading
4	WHI-SPE[15]	Serbian	2013	5,000 words	5	5	Words	-	Normal & Whispered
5	AHUMADA I,II,III[16]	Spanish	2000	-	150	250	Sentences/Digit/Word	0.5	Reading/ Extempore
6	AUS-TALK[17]	English	2012	3000 Hrs	1000		Story/Sentence/Digit/ Word		Reading/ Interview/ Story telling
7	REPERE[18]	French	2013	60 Hrs	-	-	News, Debate, TV Shows		Reading
8	ETAPE[19]	French	2011	30 Hrs	-	-	News, Debate, TV Shows	-	Reading
9	ESTER-I,II[20]	French	2006, 2009	100 Hrs	-	-	News, Debate, TV Shows	1.1	Reading

These databases are widely used in forensic applications but still they have certain lacunas. Some databases are also developed in the Indian scenario for forensic use. One database is developed by Center for Forensic Science Laboratory (CFSL), Chandigarh. Another database was developed by H. A. Patil et al., to design a robust language independent speaker recognition system. One more effort has been done by Haris et al.[1], to develop a database in English and several Indian languages of about 200 speakers in office as well as noisy places like laboratory, classroom, hostel etc. over five different channels in parallel.

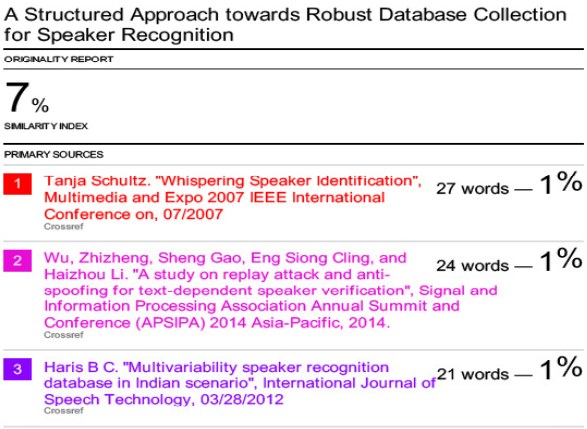
4. Conclusion and Future Work

A robust speech database can be developed considering the important factors presented in this paper like recording environment, instrument, language and many more. A database developed in this manner can be of great help in designing an efficient speaker recognition system. The system can be designed for compensating as many above explained factor as possible for better recognition rate. This work can be further extended because of the fact that, beside all these efforts and researches, still many problems faced by forensic scientists are unsolved like databases for crying speech, yelling speech, distance from microphone, age variability, and spoofing are not available along with many other challenges faced by FSR.

5. References

1. Haris BC, Pradhan G, Misra A, Prasanna SRM, Das RK, Sinha R. Multivariability Speaker Recoginition Database in Indian Scenario. IJST. 2011.
2. Ramos D, Gonzalez-Dominguez J, Arevalo E, Gonzalez-Rodriguez J. High Performance Session Variability Compensation in Forensic Automatic Speaker Recognition. Special Session on FVCFA. 2010.
3. Kenny P, Boulianne G, Ouellet P, Dumouchel P. Speaker and session variability in GMM-based Speaker Verification. IEEE Trans on ASLP. 2007;1448–60.
4. Vogt R, Sridharan S. Explicit Modelling of session Variability for Speaker Verification.CSL. 2008; 17–38.
5. Benzeghiba M, Mori RD, Deroo O, Dupont S, Erves T, Jouvet D, et al. Automatic speech recognition and speech variability: A review. Speech Communication. 2007; 763–86. https://doi.org/10.1016/j.specom.2007.02.006
6. Ichino M, Komatsu N, Jian-Gang W, Yun YW. Speaker Gender Recognition Using Score Level Fusion by AdaBoost. ICCARV. 2010. PMCid:PMC2955250
7. Haris BC, Pradhan G, Misra A, Shukla S, Sinha R, Prasanna SRM. Multi-Variability Speech Database for Robust Speaker Recognition. IEEE. 2011. https://doi.org/10.1109/NCC.2011.5734775
8. Haris BC, Sinha R. Exploring Sparse Representation Classification for Speaker Verification in Realistic Environment. Centenary Conference-Electrical Engineering. Bangalore. 2011.
9. Kelly F, Drygajlo A, Harte N. Speaker Verification with Long-Term Ageing Data. ICB. 2012. https://doi.org/10.1109/ICB.2012.6199796
10. Hautamaki RG, Kinnunen T, Hautamaki V, Leino T, Laukkanen A. I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry. Proc Interspeech. 2013. PMCid:PMC4056430
11. Chenghui G, Heming Z, Zhi T. Speaker Identification of Whispered Speech with Perceptible Mood. JM. 2014; 9(4):553–61.
12. Ajili M, Bonastre JM, Kahn J, Rossato S, Bernard G. FABIOLE, a Speech Database For Forensic Speaker Comparison. Proc of IREC Conference. 2016.
13. Vloed DV, Bouten J, Van Leeuwen DA. FRI-FRITS: A forensic speaker recognition database and some first experiments.Odyssey 2014: The Speaker and Language Recognition Workshop. 2014; 6–13.
14. Segundo ES, Alves H, Trinidad MF. CIVIL Corpus: Voice Quality for Speaker Forensic Comparison. CILC. 2013. PMCid:PMC3694487
15. Markovic GJ. Whispered speech database: design, processing and application.InternationalConference TSD. 2013.
16. Garcia JO, Rodrkguez JG, Marrero-Aguiar V. AHUMADA: A large speech corpus in Spanish for speaker characterization and identification. Speech Communication. 2000; 3:255–64. https://doi.org/10.1016/S0167-6393(99)00081-3
17. Alghowinem S, Wagner M, Goecke R. AusTalk - The Australian Speech Database: Design Framework, Recording Experience and Localisation. CITA. 2013.
18. Galibert O, Kahn J. The first official repere evaluation. Slam@Interspeech. 2013; 43–8.
19. Gravier G, Adda G, Paulson N, Carre M, Giraudel A, Galibert O. The etape corpus for the evaluation of speech-based tv content processing in the French language. LREC. 2012. PMCid:PMC3412862
20. Galliano S, Geoffrois E, Gravier G, Bonastre JF, Mostefa D, Choukri K. Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. Proceedings of LREC. 2006. PMCid:PMC1569379

Annexure-I



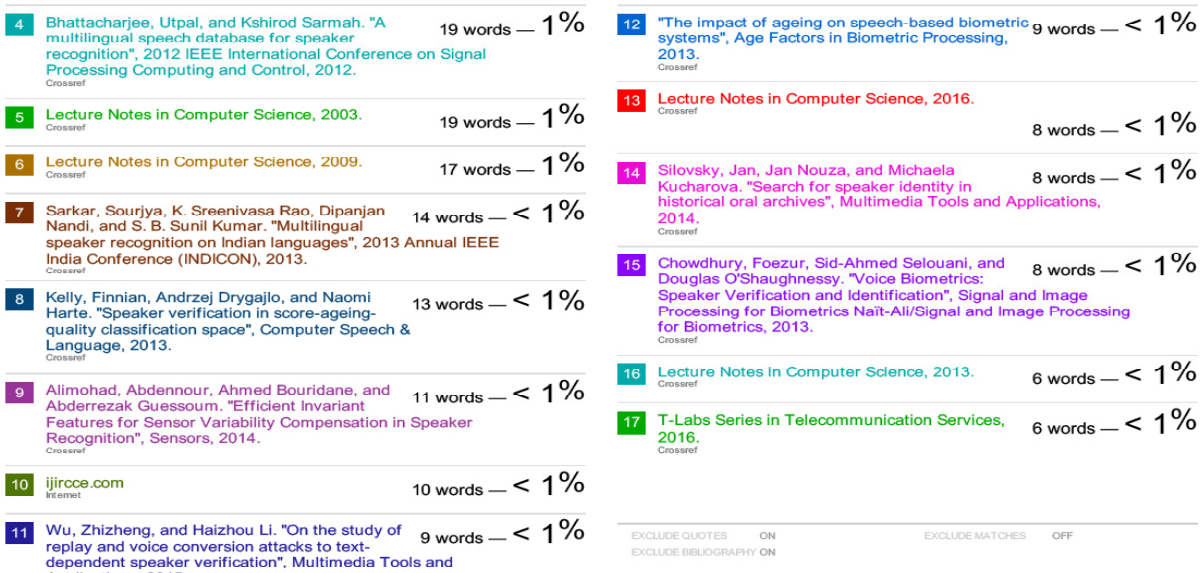
A Structured Approach towards Robust Database Collection for Speaker Recognition

ORIGINALITY REPORT

7%
SIMILARITY INDEX

PRIMARY SOURCES

1	Tanja Schultz. "Whispering Speaker Identification", Multimedia and Expo 2007 IEEE International Conference on, 07/2007 Crossref	27 words — 1%
2	Wu, Zhizheng, Sheng Gao, Eng Siong Cling, and Haizhou Li. "A study on replay attack and anti-spoofing for text-dependent speaker verification", Signal and Information Processing Association Annual Summit and Conference (APSIPA) 2014 Asia-Pacific, 2014. Crossref	24 words — 1%
3	Haris B C. "Multivariability speaker recognition database in Indian scenario", International Journal of Speech Technology, 03/28/2012 Crossref	21 words — 1%



Source: http://www.ithenticate.com/

Citation:

Pardeep Sangwan and Saurabh Bhardwaj

"**A Structured Approach towards Robust Database Collection for Speaker Recognition**",

Global Journal of Enterprise Information System. Volume-9, Issue-3, July-September, 2017. (http://informaticsjournals.com/index.php/gjeis)

DOI: 10.18311/gjeis/2017/16123

Conflict of Interest:

Author of a Paper had no conflict neither financially nor academically.