Investigating Text-independent Speaker Verification from Practically Realizable System Perspective

Rohan Kumar Das* and S. R. Mahadeva Prasanna†‡
* Department of Electrical and Computer Engineering, National University of Singapore, Singapore
† Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati, India
‡ Department Electrical Engineering, Indian Institute of Technology Dharwad, Karnataka, India
E-mail: rohankd@nus.edu.sg, prasanna@iitg.ernet.in

***Abstract*—This work projects an attempt to explore the prospects of text-independent speaker verification (SV) for practical realizable systems. Although the advancements in SV systems have gained attention towards deployable systems, the performance seems to degrade under uncontrolled conditions. A protocol for data collection is designed for the text-independent SV with student attendance as an application to create a database in a real-world scenario. The i-vector based speaker modeling is used for evaluating the performance that depicts major deviation of results from that obtained on standard database. This portrays the significance of having real-world scenario based databases for robust SV studies. Further, studies are performed related to speaker categorization, speaker confidence and model update that showcase their significance towards systems in practice. The database created in this work is available as a part of multi-style speaker recognition database.**

I. Introduction

The speech based application oriented systems have gained attention in the recent years due to the growth of research in different directions. The field of speaker verification (SV) is no exception to it and has made possible many such practical systems [1]–[3]. However, the robustness of these real-time systems depends on several factors. The systems may perform poorly for unseen condition data resulted due to acoustic variability caused by sensor noise, background noise, reverberation, etc. Similarly, the variabilities in terms of the condition of speaker, speaking rate, language, dialect are also crucial for system implementation in a real-world scenario.

The NIST speaker recognition evaluations (SRE) are held biannually that release dataset collected from a large number of speakers to evaluate the performance of SV systems [4]. These databases are considered as standard to compare different systems on a common global dataset for benchmarking system performance. However, they do not depict the real-world conditions that come into picture while deploying an SV system into user practice for some intended service. The authors of [5] have mentioned regarding the deviation of results obtained on realistic data compared to that on standard datasets. Thus, the results obtained on standard datasets may not be that conclusive for real-life applications. They have put emphasis on formulating effective solutions for data collected in a real-world scenario. This highlights the importance of research on SV to be explored on different challenging conditions, especially from the perspective of some application based systems.

A comparative assessment on different corpora and their usefulness for speaker recognition studies in the last two decades have been reported in [6]. The RATS database is one of the challenging datasets designed for SV, which is collected in a radio traffic based system environment [7]. Recently, in order to have speech data from a real-world scenario for SV research, the authors of [8] have created a database under uncontrolled conditions. This database is named as speakers in the wild (SITW) database, where the speech examples are taken from real-life interviews and casual conversations in different environments. Additionally, the speech examples of this database have background speech, laughter and many such unwanted components that make it challenging. It is observed that the performance of the state-of-the-art systems severely degrades in such conditions as reported in [9]. Thus, it depicts the significance for requirement of databases in real-world scenario for the researchers in speaker recognition community.

In this work, with the stated motivation to collect real-world data, an SV system over telephone based network is developed for the application of student attendance [10]. The system is deployed for a period of one year and a corpus having 325 speakers data with large session variation is collected in a real-world scenario. The collected data contains different variabilities in terms of health condition of speaker, mismatch in sensors, presence of different background noise in the environment, aging effect, etc. To investigate the performance on the collected data, it is evaluated by i-vector based modeling as a reference system. The regular day to day collection of data for an application system makes it challenging and hence requires investigations along different directions.

There are different works in the literature that focus on studying speaker characteristics. For instance, in [11], the authors have analyzed and categorized speakers into four different groups based on their behavior in the recognizability to an SV system. This projects towards the scope of having different strategies for dealing with different groups of speakers for a system in practice. The authors of [12] proposed a speaker confidence measure based on SNR and duration metric that is found to be useful for SV. In [13], the studies showed that session variability and template aging plays a crucial role for data collected in a longer period. Different frameworks can be adopted for compensating their effects to improve SV performance as suggested by the authors.

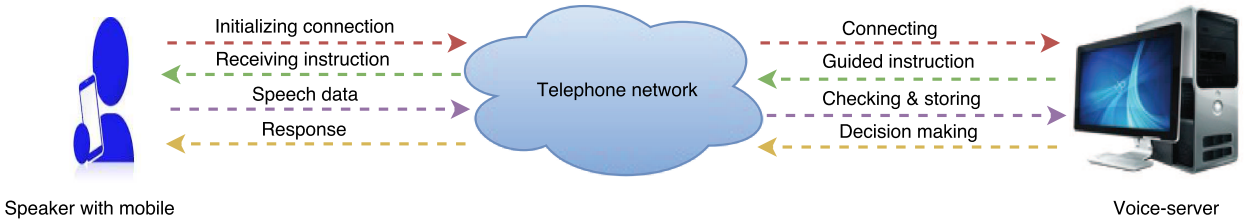

Fig. 1. Protocol for real-world data collection over telephone channel.

Along similar directions, studies have been reported in this work first to show the challenging nature of data collected in a real-world scenario. It is followed by studies related to speaker categorization, speaker confidence and model update. The contribution of this work is attributed in unveiling a text-independent SV database under real-world condition along with studies towards improving performance.

The remaining work is structured as follows. Section II mentions regarding the data collected in a real-world scenario with attendance as an application. Section III describes the details of system development and performance under standard database. In Section IV, the experimental results on the real-world data along with the strategies to improve performance are investigated and discussed. Finally, Section V provides the conclusion with future directions.

II. Database Collection Protocol in Real-world Scenario

With the widespread use of mobile phones in the current generation, the interactive voice response system (IVRS) callflow based services have gained attention for several applications. Keeping this in view, a protocol for real-world data collection has been designed over telephone channel. In this framework, an IVRS callflow guides the users for enrollment as well as testing. Figure 1 shows the overview of the SV system used for data collection under real-world scenario with attendance marking as an application. An i-vector based SV system is implemented over a voice-server that connects to the telephone network service. The users of this system have to call to a telephone number for connecting to the voice-server, where the IVRS callflow provides required instructions to the users.

The training phase of the developed system provides a four digit speaker ID to each user and asks for 3 minutes of read speech data through IVRS, which is used for speaker modeling. The speaker models are placed against each of the speaker's designated speaker ID. During testing, a speaker makes a claim against his/her speaker ID through the same IVRS callflow via testing option for verification. This system is deployed for attendance as an application for regular use of this service. The students of the department of Electronics and Electrical Engineering (EEE) at Indian Institute of Technology (IIT) Guwahati used this protocol to mark attendance from their phones for a period over a year. In this way, a database consisting of 325 speakers is collected that contains a population of 286 male and 39 female speakers. As this database is collected in an application oriented scenario, there exists a lot of variabilities. The health, emotion and many more aspects of the speakers, testing environment conditions are not the same for all the sessions. Additionally, the handsets used by the users too have a wide range that causes the handset variabilities which may have an effect on performance.

From the perspective of practically realizable text-independent SV systems, sufficient train with limited test data condition is preferred to have comfort from the users end [14]–[17]. The testing protocol for the IVRS based framework evolved across the time with user feedbacks to reduce the burden of the users from giving more amount of speech data during testing. Initially while deploying this protocol, the users are asked to produce speech for about 20 s for testing. This burdened the users to provide speech for 20 s on a daily basis, which made us modify the testing phase. The users are then asked only their roll number with name that collects speech data of about 6 s for each session. This makes the test sessions of the speakers to have different durations ranging from 6-20 s. Figure 2 shows the statistics of the duration of utterances collected from real-world data for train and the test set after performing voice activity detection (VAD). It could be observed that the utterances of the train set have speech durations more than 70 s for most of them. On the other hand, the speech duration for test data lies between 2-10 s. Thus, the database can be considered for sufficient train and limited test data based SV studies.

It is to be mentioned that the above mentioned database is collected as a part of multi-level SV system that includes three different SV modalities [18]. The three different modalities include voice-password, text-dependent and text-independent SV systems and the collective database is referred as multi-style speaker recognition database. Its details with specifications can be found in [19]. Additionally, the database is available for public use and can be availed by contacting the authors.

III. System Description

This section mentions regarding the SV system developed for conducting studies in this work. The front-end processing, i-vector based SV framework along with its results on standard dataset are described in the following subsections.

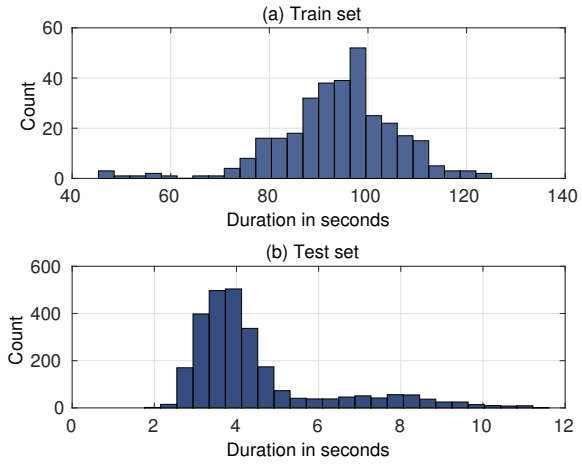

Fig. 2. Histograms for speech duration available after VAD for train and test set of data collected in real-world scenario.

A. Front-end Processing

The utterances are processed by considering them as blocks of 20 ms with a shift of 10 ms. 39-dimensional (13-base+13-Δ+13-$\Delta\Delta$) mel frequency cepstral coefficient (MFCC) features are extracted for each block of frames. An energy based VAD is applied on the utterances to detect the speech regions and the features of those portions are considered for normalization in cepstral domain by cepstral mean and variance normalization (CMVN) technique [20]. The normalized features are then used for SV studies.

B. i-vector System

The i-vector based SV is one of the recent in technologies in this area developed couple of years back [21]. This approach projects the Gaussian mixture model (GMM) mean supervectors of an utterance into a compact low dimensional space via a transformation matrix that captures all sorts of acoustic variabilities. The low rank representation vector is found to have dominant speaker-specific information useful for speaker modeling. Additionally, this kind of framework is useful for having channel/session compensation that provides an edge over other conventional modeling approaches for SV.

C. Studies on Standard Database

The recent SRE evaluations tend to focus more on robustness to different scenarios, for which different conditions are introduced. The NIST SRE 2012 database is one of the latest and largest database, where telephone as well as microphone speech is considered. It has several conditions for observing SV performance trend with and without presence of noise. There are five conditions altogether for the core-core set of tasks for SRE 2012 including telephone and microphone channel database.

The i-vector system is developed on the NIST SRE 2012 database. Two gender dependent universal background models (UBM) of 1024 mixture components are trained using background data. Then sufficient statistics of each utterance are extracted for male and female set using respective UBM. The statistics of the background data are used to learn the total variability matrix of 800 dimension. Linear discriminant analysis (LDA) and within class covariance normalization (WCCN) are used for channel/session compensation [22], [23]. It is to be noted that 250-dimensional LDA and full dimensional WCCN matrices are used in this work. The verification of a trial is done according to the protocol of NIST SRE 2012 evaluation plan. The results are reported in terms of equal error rate (EER) and detection cost function (DCF).

TABLE I
PERFORMANCE ON EVALUATION SET OF NIST SRE 2012.

Evaluation Condition		actDCF	minDCF	EER (%)
Phone call	No noise	0.4629	0.3855	3.63
	Added noise	0.7461	0.6856	6.94
	Noisy env.	0.4985	0.4274	3.48
Interview	No noise	0.6347	0.5681	7.35
	Added noise	0.5593	0.4663	5.27

TABLE II
PERFORMANCE ON REAL-WORLD DATABASE.

Evaluation Condition	EER (%)
Phone call	11.50

Table I shows the performance of the developed i-vector system on the evaluation set of NIST SRE 2012 dataset, which depicts that it is able to handle noisy conditions provided for the evaluation. This shows the potential of this framework for practically realizable systems. However, the noisy conditions of this database are simulated for evaluating system performance in such scenarios. Therefore, it becomes interesting to explore the significance of the state-of-the-art SV systems in handling real-world data with wide range of variabilities that is collected on a regular basis from some deployable system.

IV. Experimental Results and Discussion

This section reports the SV studies on the database collected from real-world scenario. It is followed by exploration of few directions towards improving performance in such conditions.

A. Studies on Real-world Database

The database collected from a student attendance system as explained in Section II is used for the studies under real-world scenario. As discussed earlier the database is collected over a period of one year for attendance application and has a wide range of variabilities involved. The i-vector based SV system described in the previous section is used to evaluate the system performance. Table II shows the performance on this database which reveals that there is a deviation of results from that obtained on the standard database. Additionally, it is to be mentioned that as the database is collected as a part of student attendance system, the students are informed at the end of testing whether their claim has successfully authenticated for marking attendance. For this, initially in house trials are made by a closed set of speakers to learn the threshold that can discriminate the genuine and impostor trials. In this system,

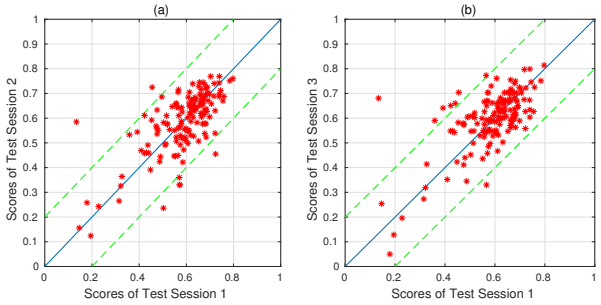

Fig. 3. Scatter plot of scores from different test sessions: (a) Session 1 vs. Session 2 (b) Session 1 vs. Session 3.

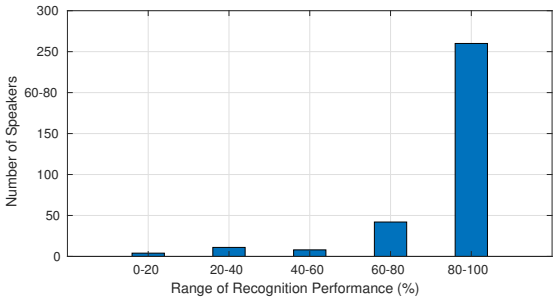

Fig. 4. Recognition performance for speaker population.

the cosine score threshold is obtained at 0.5, which is used for decision over the telephone based system described in Section II.

The deviation of results obtained in the case of real-world database from that obtained for standard database motivated for exploring the former. It is found that most of the utterances have background speaker's speech, different kinds of noise due to sensor, background and many such external factors that may have resulted in poor speaker-specific characterization. Next we study few directions that can be explored for text-independent SV systems in real-world practice.

B. Speaker Categorization

The studies in [11] show that there exists different kinds of speaker groups in a population. The authors categorized four different group of speakers namely, sheep, goat, lamb and wolf based on their behavior to the SV system. In this work, we collected the database as described in Section II in real-world scenario with a sizeable population with large number of sessions. In order to characterize the speakers based on their behavior we have observed the scores produced during attendance marking.

Figure 3 shows the scatter plots of scores against same speaker models from two sessions. Three different test sessions of 150 speakers from the collected database are considered to illustrate their behavior. Out of this three sessions, one session is kept common for both the subplots of Figure 3 so that the speaker trend can be visualized. The center diagonal line is used for reference and the two dotted lines on both sides are considered as tolerance bands of ± 0.2. Under ideal condition, the scores plotted should follow the reference diagonal line and should lie between the tolerance bands to convey their stable trend. It is observed that scores from most of the speakers occupy the region between the tolerance bands. Further, there are majority of speakers producing cosine scores higher than 0.5 that projects those speakers, which can be identified well by the system. On the other hand, there are few speakers that have consistently low scores as can be observed from Figure 3.

Based on the observations the speakers are categorized into two categories. The first one is referred to as *positive speakers*, which contains the speakers producing higher scores against like speaker models that are well classified by the system. On the contrary, the other class is named as *negative speakers* that belongs to the speakers which are not often correctly detected by the system. It is to be mentioned that in any text-independent SV system in practice this kind of categories are expected to present in the speaker population. Thus, it may be useful to identify them and then to apply different strategies for dealing with them from practical realizable system perspective.

C. Speaker Confidence and Model Update

The text-independent systems in application scenario have regular testings for speaker authentication. Session variability is found to be a crucial factor for SV as mentioned in the literature. Further, from the studies performed related to speaker characterization shows that the behavior of speakers are different to the SV systems. In this regard, it can be useful to follow the statistics of historical trials of a speaker. A subset of test set with former test sessions of the collected database is considered to study the trend in speaker's historical data. Figure 4 shows the analysis on recognition performance of the trials against same speaker models. The performance is grouped in five different categories as shown in the figure with five ranges. It is observed that the majority of the speakers have higher recognition accuracy, however the ones with less performance is our concern in case of a practical system. The speakers who continuously perform poorly may require careful attention to investigate the fault behind their rejection. A measure of confidence can be used from the performance of speaker's past historical trials while evaluating the subsequent trials in a deployable system.

In this work, to utilize the historical data of the speakers, the speaker models are retrained including the first 3 test sessions. The retrained models are then used for verification of subsequent trials. Figure 5 shows the scores obtained from three subsequent test sessions against same speaker models with original and their retrained models. It can be noticed that the scores obtained against retrained model are comparatively higher than that against the original model. This showcases the significance of speaker model update from the historical data for a practical system. Further, it is to be noted that we have not parameterized a speaker confidence measure to have an update in speaker model. The models are updated considering their first 3 test sessions irrespective of the *positive*

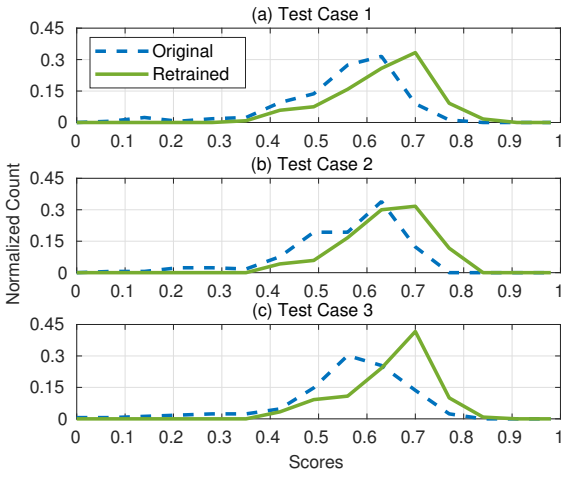

Fig. 5. Histogram of scores obtained for test trials of three different sessions against original and retrained speaker model.

or *negative speaker* class mentioned previously. The use of retrained models in the database reduces the EER from 11.50% to 8.21% indicating its potential for realization of practical text-independent SV systems.

V. Conclusion

This work investigates text-independent SV in real-world scenario for application oriented systems. It has been found that the there lies a gap between the results on laboratory based experiments and real-world setup. In this regard, a text-independent SV system is developed using i-vector based speaker modeling over telephone network and then deployed for student attendance application. A database comprising 325 speakers is collected from this real-world scenario that is made available as a part of multi-style speaker recognition database for public use. The experiments are conducted to show the challenging nature of this real-world data by comparing the results with standard database. Further, studies are performed on the real-world data that have scope from the perspective of practical realizable systems. These studies include speaker characterization, speaker confidence and model update that show their importance towards utilizing them for systems in practice. The future work will focus on parameterizing speaker confidence for model update and related studies to increase the potential of text-independent SV in a real-world scenario.

VI. Acknowledgement

The authors would like to acknowledge all the project members and the student volunteers of IIT Guwahati that have been a part of the creation of the corpus discussed in this work. The work of the first author is supported by Programmatic Grant No. A1687b0033 from the Singapore government's Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain).

References

[1] K.-A. Lee, A. Larcher, H. Thai, B. Ma, and H. Li, "Joint application of speech and speaker recognition for automation and security in smart home," in *INTERSPEECH*, 2011, pp. 3317–3318.

[2] D. Chakrabarty, S. R. Mahadeva Prasanna, and R. K. Das, "Development and evaluation of online text-independent speaker verification system for remote person authentication," *International Journal of Speech Technology*, vol. 16, no. 1, pp. 75–88, 2013.

[3] R. Ramos-Lara, M. Lpez-Garca, E. Cant-Navarro, and L. Puente-Rodriguez, "Real-time speaker verification system implemented on reconfigurable hardware," *Journal of Signal Processing Systems*, vol. 71, no. 2, pp. 89–103, 2013.

[4] NIST speaker recognition evaluations. [Online]. Available: www.nist.gov/itl/iad/mig/speaker-recognition

[5] J. H. Hansen and H. Boil, "Robustness in speech, speaker, and language recognition: youve got to know your limitations," in *Interspeech 2016, San Francisco*, 2016, pp. 2766–2770.

[6] D. E. Sturim, P. A. Torres-Carrasquillo, and J. P. Campbell, "Corpora for the evaluation of robust speaker recognition systems," in *Interspeech 2016, San Francisco*, 2016, pp. 2776–2780.

[7] K. Walker and S. Strassel, "The RATS radio traffic collection system," in *Odyssey 2012: The Speaker and Language Recognition Workshop, Singapore, June 25-28, 2012*, 2012, pp. 291–297.

[8] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (sitw) speaker recognition database," in *Interspeech 2016, San Francisco*, 2016, pp. 818–822.

[9] ——, "The 2016 speakers in the wild speaker recognition evaluation," in *Interspeech 2016, San Francisco*, 2016, pp. 823–827.

[10] S. Dey, S. Barman, R. K. Bhukya, R. K. Das, Haris B C, S. R. M. Prasanna, and R. Sinha, "Speech biometric based attendance system," in *National Conference on Communications (NCC) 2014, IIT Kanpur*, 2014.

[11] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "Sheep, goats, lambs and wolves a statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation," in *International Conference on Spoken Language Processing*, 1998.

[12] M. C. Huggins and J. J. Grieco, "Confidence metrics for speaker identification," in *INTERSPEECH 2002, Denver, Colorado, USA*, Sep 2002.

[13] Rohan Kumar Das, S. Jelil, and S. R. M. Prasanna, "Exploring session variability and template aging in speaker verification for fixed phrase short utterances," in *Interspeech 2016*, 2016, pp. 445–449.

[14] Rohan Kumar Das and S. R. M. Prasanna, *Speaker Verification for Variable Duration Segments and the Effect of Session Variability*. Lecture Notes in Electrical Engineering: Springer, 2015, ch. 16, pp. 193–200.

[15] Rohan Kumar Das and S. R. M. Prasanna, "Speaker verification from short utterance perspective: A review," *IETE Technical Review*, pp. 1–19, 2017.

[16] Rohan Kumar Das and S. R. M. Prasanna, "Exploring different attributes of source information for speaker verification with limited test data," *The Journal of the Acoustical Society of America*, vol. 140, no. 1, pp. 184–190, 2016.

[17] Rohan Kumar Das, A. B. Manam, and S. R. M. Prasanna, "Exploring kernel discriminant analysis for speaker verification with limited test data," *Pattern Recognition Letters*, vol. 98, pp. 26 – 31, 2017.

[18] Rohan Kumar Das, S. Jelil, and S. R. M. Prasanna, "Development of multi-level speech based person authentication system," *Journal of Signal Processing Systems*, vol. 88, no. 3, pp. 259–271, Sep 2017.

[19] ——, "Multi-style speaker recognition database in practical conditions," *International Journal of Speech Technology*, Nov 2017.

[20] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, Apr 1981.

[21] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[22] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley and Sons, 2000.

[23] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2006.