Text-independent Speaker Verification with Limited Test Data from the Perspective of Practical Systems

Rohan Kumar Das and S. R. Mahadeva Prasanna

Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati, Guwahati-781039
{rohankd, prasanna}@iitg.ernet.in

Abstract

This work is an attempt to take up speaker verification for deployable systems. In this regard a framework with sufficient train and limited test data (preferably $\leq$ 10 sec) is formulated that is feasible in practical system perspective with respect to user comfort and effective decision delivery. The drawbacks of dealing with limited test data is compensated with a framework having different source features, text-constrained model and effective pattern recognition approaches. Some studies with respect to practical system based on session variability and template aging are also included under this work. Utilizing all these exploration a proposal of an improved framework for handling limited test data for practical system perspective has been made.

Index Terms: speaker verification, limited data, source features, text-constrained, session variability

1. Introduction

Over the time, research on different fields has progressed by many folds and has witnessed many breakthroughs. Speaker verification (SV) is no exception to it, that projects scope towards field deployable systems for regular use. In this direction, some attempts have been already made by different researchers across the globe that shows potential in application services [1, 2, 3, 4, 5]. Nevertheless there comes many hurdles on its way for deployment, one such kind is amount of speech data available for verification of a claim. From user comfort and effective decision making less amount of speech data is expected. But as the speech available for testing reduces, it affects the performance of the SV systems significantly. The efficacy of the current state-of-the-art i-vector based speaker modeling deviates with less amount of speech data [6, 7].

Motivated by the drawbacks of SV system performance under limited data, this work focuses on a framework to handle the limited data scenario for SV. An architecture with sufficient train and limited test data is proposed as a suitable approach to overcome this hurdle of limited data, as one time sufficient amount of data may be collected from the users for a practical system in a cooperative scenario. Whereas test data of short durations are expected to reduce the burden from the intended users for delivering more amount of speech on regular use. Limited test speech segments of $\leq$ 10 sec is considered for the experimental studies for this SV study. Further this framework of text-independent SV with limited test data is explored for possible improvements by use of complimentary features, different strategies and pattern recognition techniques. Finally in addition to all these, an improved framework is expected to handle limited test data SV in a text-independent framework having scope towards deployable systems. Also different scenarios and conditions that effects SV performance are also important aspect of this work.

The rest of the work can be seen as: Section 2 describes the importance of source information and highlights the different attributes of source information that is useful for SV with limited test data. In Section 3, a framework of text-constrained model is mentioned that is found to have a significant impact on SV performance. Some practical system based studies on session variability and template aging are explored in Section 4. Section 5 shows the future work plan with overall framework to handle limited test data in a text-independent scenario.

2. Importance of Source Information and their Different Attributes for SV

The voice source features represent the excitation source nature which can be used for discriminating speakers as each speaker has different structure of glottis and muscle structure. However these features are not as much effective as the conventional vocal tract features. In spite of this, they are found to work well under limited data scenario as their dependence is very less on phonetic content unlike the vocal tract features.

The linear prediction (LP) residual that contain the excitation source information has a noise like structure. This makes the information captured by it less discriminative for speaker-specific knowledge. Three different source features namely mel power difference in subband spectrum (MPDSS) [8], residual mel frequency cepstral coefficients (RMFCC) [9] and discrete cosine transform of linear prediction residual (DCTILPR) [10, 11] which involve processing of LP residual signal in spectral, cepstral and temporal domain, respectively. The three source features MPDSS, RMFCC and DCTILPR are found to carry different attributes of source information namely, periodicity, smoothed spectrum information and shape of the glottal signal, respectively from the studies made in [12]. The fusion of these features is carried at the score level in an i-vector based SV framework, which outperformed the baseline framework based results using conventional mel frequency cepstral coefficients (MFCC) features. Also the source features in fusion to MFCC yielded a large improvement over the baseline system. This projected the importance of source features for limited test data framework along with their different attributes that on fusion can help in improving SV performance.

3. Significance of Text-constrained Models for Limited Data Speaker Verification

The text-dependent SV modality is based on same lexical content during training and testing from all the users. This pro-

vides an edge over the text-independent framework and thus the former deals with short fixed phrases of 3-4 seconds of duration. The same reflected for a better match between the train and test sessions, when Gaussian posteriorgram based SV framework is used for text-dependent modality with creation of text-specific and speaker-specific Gaussian mixture model (GMM) [13]. In order to replicate the advantage of lexical match in text-independent modality, an SV system architecture based on text-constrained model is proposed where the speech examples of speakers are made with a user-specific phrase and a fixed phrase together around 10 sec duration which is same during training and testing [14]. In this manner the lexical match between train and the test session is made in a supervised way along with maintaining text-independent nature as the user-specific phrase is involved. Also some experimental studies are made to highlight the importance of phonetic match between train and test sessions, the details of which can be seen from [14]. The immediate future work of this is to plan the text-constrained model based framework in an unsupervised manner without putting any restrictions on the users of producing a user-specific phrase along with a common fixed phrase utterance.

4. Effect of Session Variability and Template Aging

There comes several issues while having a practical deployable systems. Session variability and template aging are two of them that we have considered under this work [15]. Recently made available RedDots database has put forward different insightful directions to research in the field of SV[16]. This database is collected over a year period that involves a large number of sessions from 62 speakers. However it is having mainly a text-dependent modality based organization with a text-prompted enrollment condition, that is similar to long enrollment sessions containing multiple fixed phrases and a fixed phrase for testing, that has resemblance to the framework considered in this work. In a practical system point of view the effect of session variability and template aging has definite significance. A framework for handling session variability in an implicit manner is proposed by considering first, middle and last session of the speakers for modeling and remaining sessions for testing [15]. This framework produced better results than the baseline framework that depicts the impact of session variability in deployable scenario. Also the experiments related to template aging with modeling the first three and the last there sessions show that the speaker models are relatively more robust with the use of later sessions for modeling [15]. This is due to the fact that the speakers get acquainted with the system as they regularly provide data. Also it is assumed that some vast speaker characteristics may have evolved over time that produced improved results by modeling speakers with last three sessions.

5. Future Work Plan

The future work plan is to have some robust pattern recognition techniques from the perspective of handling limited test data conditions to provide better discrimination across speakers. Then to propose an overall framework that will consider the different source features in fusion to conventional MFCC based vocal tract feature with robust pattern recognition techniques along with exploitation of a text-constrained framework, which may overcome the hurdles present in using limited test data in text-independent SV from the perspective of practical deployable systems.

6. 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] 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*, 2014.

[4] 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.

[5] Rohan Kumar Das, S. Jelil, and S. R. Mahadeva Prasanna, "Development of multi-level speech based person authentication system," *Journal of Signal Processing Systems*, pp. 1–13, 2016. [Online]. Available: http://dx.doi.org/10.1007/s11265-016-1148-z

[6] 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.

[7] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, , and M. Mason, "i-vector based speaker recognition on short utterances," in *Interspeech 2011*, 2011.

[8] Rohan Kumar Das, Debadatta Pati, and S. R. M. Prasanna, "Different aspects of source information for limited data speaker verification," in *National Conference on Communications (NCC) 2015*, 2015.

[9] D. Pati and S. R. M. Prasanna, "Speaker information from subband energies of linear prediction residual," in *National Conference on Communications (NCC), 2010*, Jan 2010, pp. 1–4.

[10] A. G. Ramakrishnan, B. Abhiram, and S. R. M. Prasanna, "Voice source characterization using pitch synchronous discrete cosine transform for speaker identification," *JASA Express Letters*, vol. 137, pp. EL469–EL475, 2015.

[11] Rohan Kumar Das, Abhiram B., S. R. M. Prasanna, and A. G. Ramakrishnan, "Combining source and system information for limited data speaker verification," in *Interspeech 2014, Singapore*, 2014, pp. 1836–1840.

[12] Rohan Kumar Das and S. R. Mahadeva 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.

[13] Sarfaraz Jelil, Rohan Kumar Das, Rohit Sinha, and S. R. M. Prasanna, "Speaker verification using gaussian posteriorgrams on fixed phrase short utterances," in *Interspeech 2015, Dresden, Germany*, 2015, pp. 1042–1046.

[14] Rohan Kumar Das, Sarfaraz Jelil, and S. R. M. Prasanna, "Significance of constraining text in limited data text-independent speaker verification," in *International Conference on Signal Processing and Communications (SPCOM) 2016*, 2016.

[15] Rohan Kumar Das, Sarfaraz Jelil, and S. R. Mahadeva Prasanna, "Exploring session variability and template aging in speaker verification for fixed phrase short utterances," in *Interspeech 2016, San Francisco*, 2016.

[16] K. A. Lee, A. Larcher, W. Guangsen, K. Patrick, N. Brummer, D. van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma, H. Li, T. Stafylakis, J. Alam, A. Swart, and J. Perez, "The RedDots data collection for speaker recognition," in *Interspeech 2015, Dresden, Germany*, 2015, pp. 2996–3000.