Incorporating Source Features, Acoustic-phonetic Information and Suitable Pattern Recognition Approach for Limited Test Data Speaker Verification

Rohan Kumar Das

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

Abstract

This work concentrates on extending the scope of text-independent speaker verification (SV) systems for practical application oriented services at par with the existing other biometric measures. In this direction, sufficient train with limited test data ($\leq$ 10 s) based framework is considered for having user comfortness along with efficient decision delivery. Different excitation source features, acoustic-phonetic information and suitable pattern recognition approaches are investigated to obtain improved performance in such a scenario. Finally, the stated directions are fit to a common architecture to be benefited from each of them. The final combined SV framework shows the potential for field deployable systems providing an improved performance. Additionally, some of the issues that come across the development of practical SV systems are also explored with possible suggestions to deal with them.

Index Terms: speaker verification, source features, acoustic-phonetic information, kernel discriminant analysis

1. Introduction

The exponential growth in research has led to systems into deployment and the field of speaker verification (SV) too has experienced the same. In the last decade several application oriented systems based on SV technology has emerged out that have gained attention showing prospects towards systems into practice [1, 2, 3, 4, 5]. While considering systems for practical deployment, the amount of speech data is expected to be minimal in order to provide user comfort and effective decision delivery. However, in case of text-independent SV, the amount of speech data available plays a major role in SV performance. It is found that the performance achieved with state-of-the-art i-vector based speaker modeling degrades as the amount of speech data is reduced [6, 7, 8]. Thus, limited data for SV becomes a critical point, which motivated to choose the problem statement of the thesis on this ground.

The significance and motivation of using limited data made to come up with a framework of sufficient train with limited test data for a practical application oriented system. This is based on the fact that one time sufficient data can be taken from the users for enrollment, while the testings for regular usage of the system has to be of short duration. For studies, short segments of duration $\leq$ 10 s is considered as limited data used for testing. The aim of the current work is to achieve an improved performance in a scenario using limited test data for SV. In order to accomplishing this goal, different features, acoustic-phonetic information and suitable pattern recognition techniques are explored, which provides the foundation of this work. Additionally, some of the issues that occur in practical systems are investigated that are under the scope of this work.

The remaining organization of this work can be seen as: Section 2 mentions regarding the different attributes of source information that is useful for SV with limited test data. The exploitations made with respect to acoustic-phonetic information is detailed in Section 3. Section 4 describes the kernel based discriminant analysis useful for limited data perspective. The combined framework developed using the stated exploration is explained in Section 5. In Section 6, some issues related to the practical systems are investigated. The future work plan followed by summary and contributions are mentioned in Section 7 and Section 8, respectively.

2. Excitation Source Features Having Different Attributes

In case of limited data based SV, as there is a very small amount of data available for characterization of speakers, there comes a need to have alternative features which may be useful to capture speaker characteristics. Although, mel frequency cepstral coefficient (MFCC) features are ubiquitous in the field of speech processing, there are features which can have speaker information. The excitation source features stand as one of those kind and their importance has been demonstrated for SV [9, 10, 11]. Even though, these voice source features are not as much effective as the conventional vocal tract features, their fusion with vocal tract features has always been advantageous [9, 11]. Thus the source features are taken as a step to explore for improving performance in a limited test data based SV system.

The noise like structure of linear prediction (LP) residual makes the source information less discriminative to capture speaker-specific information. It is hypothesized that considering source features of different attributes can lead towards improved speaker characterization. In this direction, three source features mel power difference in subband spectrum (MPDSS) [12], residual mel frequency cepstral coefficients (RMFCC) [13] and discrete cosine transform of linear prediction residual (DCTILPR) [14, 15] are explored from their origin to investigate the nature of speaker characteristics carried by each of them. The explorations depict that each of the considered source feature possesses unique attribute of excitation source, given by periodicity, smoothed spectrum information and shape of the glottal signal, respectively [16]. Although, LP residual is the common signal used to extract these features, due to different signal processing mechanisms involved and evidence captured, each of them represent specific aspect of excitation source. The three source features are then fused as mentioned in [16], which provides improved SV performance. Further, their fusion with MFCC features enhances the performance by large margin indicating their significance for SV with limited test data.

3. Significance of Acoustic-phonetic Information

The acoustic-phonetic information in an utterance spoken by the speaker has a definite significance in SV. The match in lexical content can be utilized for SV, which comes into the picture from the genesis of text-dependent mode of SV. This became further strengthened when Gaussian posteriorgram based features are extracted using text-specific and sentence-specific Gaussian mixture model (GMM) that produced a better match between the train and test sessions [17]. However, in text-independent SV, there is no restriction on what to be spoken during training as well as testing. Thus, it becomes challenging to utilize the acoustic-phonetic information in a text-independent SV scenario. In this direction, a text-constraint model based approach is proposed to have a match of lexical contents between train and test sessions for a text-independent SV scenario in an explicit manner [18, 19]. Further, in order to capture the acoustic-phonetic information in an implicit manner vocal-tract constriction (VTC) evidence is used , which capture the nature of constriction while producing different phonetic units [20]. The use of VTC evidence along with the MFCC features results into an improvement that indicates the benefit of utilizing acoustic-phonetic information in SV.

4. Kernel based Discriminant Analysis

The contributions towards having improved performance after exploring different attributes of source information and then with utilization of speaker-specific acoustic-phonetic information, another direction is investigated. The observations from i-vector based speaker modeling shows that the i-vectors of the short utterances vary much due to large variation in the phonetic content in each case [21]. Therefore, the conventional pattern recognition techniques for channel/session compensation could not distinctly separate the i-vectors across different classes. In this regard, kernel discriminant analysis (KDA) is explored which transforms the feature vectors into a higher dimensional space and then performs discriminant analysis [22]. The KDA when used at the back-end of i-vector based speaker modeling for channel/session compensation performs better than the other existing techniques like linear discriminant analysis (LDA) followed by within class covariance normalization (WCCN) and probabilistic linear discriminant analysis (PLDA). Further, it is observed that the KDA based framework is more useful while dealing with the limited test data based scenario.

5. Combined Framework

The work then focuses on bringing out a framework combining all the explorations made for having improved speaker characterization for SV with limited test data. In this final framework, VTC feature is extracted and combined with conventional MFCC features at the feature level and i-vector based modeling is done for speaker representation. Three parallel systems are build using the different attributes of excitation source information based on source features MPDSS, RMFCC and DCTILPR. The KDA based channel/session compensation is applied at the back-end on each of the systems developed with the stated features. Finally, a score level fusion is made for all the systems to complete the combined framework with the different explorations made. The developed combined system is able to handle SV with limited test data to a large extent enhancing the baseline system performance [23].

6. Investigating Issues in Practical Scenario

In this work, few issues that come in a practical system are investigated, that include mismatch in speech tempo, session variability and template aging. It has been found in the literature that the mismatch in train and test conditions degrades the SV performance. The mismatch in speaking rate between train and test sessions is one of those. A prosody modification method is employed to modify the speaking rate of the test speech according to that to the train speech to compensate the mismatch in speech tempo [24, 25]. For investigating the session variability and template aging, the RedDots database designed for short utterance studies is used [26]. It is collected over a period of one year and hence has a lot of session variability in the trials collected from the users. It is observed from the studies that considering the first, middle and last sessions of speakers can provide improved results than that obtained with the baseline framework [27]. Further, the studies related to template aging depict that the later sessions of the speaker are more robust for capturing speaker information. Thus these project that there is a need to update the speaker models in regular intervals in case of a practical system.

7. Future Work

The future work is motivated by the current results that have been discussed. Some of the future works may be seen as,

- Explore the importance of the excitation source features within the glottal region for more robustness.
- To have evidence like vowel roundness, frontness for detailed capture of acoustic-phonetic information.
- Identify specific variabilities for limited data SV and implement specific compensation methods to deal with.
- From the view of practical system, source features in degraded condition, spoofing attacks may be explored.

8. Summary and Contributions

This work focuses on developing an improved SV framework using three different directions. These directions include excitation source features, acoustic-phonetic information and suitable pattern recognition approach. Further, some of the issues of practical SV systems are also investigated alongwith possible solutions for them. The prime contributions can be observed as:

1. Bringing out an SV framework having sufficient train with limited test data for application oriented systems.
2. Exploring different attributes of source information and significance for SV with limited test data on fusion.
3. Proposal of a text-constraint model based framework and explicit/implicit utilization of speaker-specific acoustic-phonetic information for speaker modeling.
4. Exploring kernel based discriminant analysis and its scope for SV with limited test data.
5. A combined framework involving the stated three directions for dealing with limited test data.
6. Investigating some issues in practical systems: mismatch speech tempo, session variability and template aging.

9. Acknowledgement

The work is incomplete without the simulating and regular discussions with supervisor Prof. S. R. Mahadeva Prasanna, whom I would like to acknowledge for the thesis work.

10. 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 (NCC) 2014, IIT Kanpur*, 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 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.

[9] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 52–55, Jan 2006.

[10] J. Gudnason and M. Brookes, "Voice source cepstrum coefficients for speaker identification," in *Proc. ICASSP*, 2008, pp. 4821–4824.

[11] S. R. M. Prasanna, C. Gupta, and B. Yegananarayana, "Extraction of speaker specific information from linear prediction residual of speech," *Speech Communication*, vol. 48, pp. 1243–1261, 2006.

[12] 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, IIT Bombay*, 2015.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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, IISc Bangalore*, 2016.

[19] ——, "Exploring text-constraint models and source information for long-enrollment with short-test speaker verification," *Circuits, Systems and Signal Processing, Springer, (Under Review)*.

[20] B. D. Sarma and S. R. M. Prasanna, "Analysis of vocal tract constrictions using zero frequency filtering," *IEEE Signal Processing Letters*, vol. 21, no. 12, pp. 1481–1485, Dec 2014.

[21] A. Kanagasundaram, D. Dean, S. Sridharan, J. Gonzalez-Dominguez, J. Gonzalez-Rodriguez, and D. Ramos, "Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques," *Speech Communication*, vol. 59, pp. 69 – 82, 2014.

[22] 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, (Under Review)*.

[23] Rohan Kumar Das and S. R. M. Prasanna, "Investigating a framework for text-independent speaker verification systems with limited test data," *IEEE Transactions on Information Forensics and Security, (Under Review)*.

[24] B. Sharma and S. R. M. Prasanna, "Faster prosody modification using time scaling of epochs," in *Annual IEEE India Conference (INDICON) 2014*, 2014, pp. 1–5.

[25] Rohan Kumar Das, Bidisha Sharma, and S. R. M. Prasanna, "Significance of duration modification for speaker verification under mismatch speech tempo condition," *International Journal of Speech Technology, Springer, (Under Review)*.

[26] 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.

[27] 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.