INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Exploring Session Variability and Template Aging in Speaker Verification for Fixed Phrase Short Utterances

Rohan Kumar Das, Sarfaraz Jelil and S. R. Mahadeva Prasanna

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

Abstract

This work highlights the impact of session variability and template aging on speaker verification (SV) using fixed phrase short utterances from the RedDots database. These have been collected over a period of one year and contain a large number of sessions per speaker. Session variation has been found to have a direct influence on SV performance and its significance is even greater for the case of fixed phrase short utterances as a very small amount of speech data is involved for speaker modeling as well as testing. Similarly for a practical deployable SV system when there is large session variation involved over a period of time, the template aging of the speakers may effect the SV performance. This work attempts to address some issues related to session variability and template aging of speakers which are found for data having large session variability, that if considered can be utilized for improving the performance of an SV system.

Index Terms: speaker verification, session variability, template aging

1. Introduction

The current achievements in the field of speaker verification (SV) have found wide spread use in various application oriented services. These application oriented services mainly focus on short utterances for recognizing speakers due to the constraint of time involved which can provide feasibility in deployment. Fixed phrase short utterances provide the basis of the short utterance case as less amount of time is involved during training and testing. However when we go for deployable systems with regular use and for a long period of time, the effect of session variability and template aging may reflect some degradation in recognition performance.

Several works have been done in the past to address the issue of session variability. In [1], the authors explicitly model session variability by generating a session dependent factor in a low dimensional subspace. The efficacy of this approach is proved for the NIST database in a text-independent framework, which clearly showed the significance of session variation for SV performance. Another way to handle this session variation is to have session compensation techniques to reduce the effect of session variability. There are different approaches for session compensation, some of which are joint factor analysis (JFA), linear discriminant analysis (LDA), nuisance attribute projection (NAP) etc. [2, 3, 4]. These approaches are found to help SV performance by providing session compensation. The authors in [5] have made a comparison of different session variability compensation approaches in SV. The work reported in [6] proposes an approach based maximum-likelihood linear regression (MLLR) adaptation that transforms for multiple recognition models and phone classes for session variability normalization which improves the SV performance. Thus, the impact of session variation is found to be very crucial for SV performance.

The aging phenomenon in different biometrics has been an interesting aspect for dealing with cutting edge technologies from a practical deployable system point of view [7]. Considering speech biometric based systems, the aging effect of speaker models has not been addressed to a large extent. The studies of [8] carried out on 22 speakers data collected for three sessions with 1-2 months of gap show that time lapse in test session degrades the performance to an extent. In [9], the authors have made studies on long term aging data over 18 speakers for 30-60 years span that show the genuine scores of speakers are affected severely than that of the impostor scores with the aging of the speaker templates. The work in [10] reports that the error rate doubles when the train and the test sessions have an interval of more than a month. In [11], the author conducts a study for exploring the aging effect for data collected for an interval of four years and reports the amount of degradation in performance is gradually more for the trials having larger time interval from training.

The limited exploration in the area of template aging is mainly due to the lack of availability in databases having large session variation from a sizeable population of speakers. The recently made available data as a part of RedDots project has opened the doors towards exploring template aging for fixed phrase short utterances [12]. In this current work, the effect of session variability is addressed by the creation of a speaker model with session variated three templates (first, middle and last sessions) and then testing by remaining templates of the RedDots dataset. This framework for creation of speaker models by data having large session variation is expected to perform better than that of the baseline due to consideration of the session variability for speaker modeling. Further, template aging studies are conducted with creation of speaker models with two approaches, where the first one is based on creation of speaker models with first three sessions and the latter is using the last three sessions. It is hypothesized that there may be a significant difference in speaker characteristics from first three sessions to last three session that is collected over a span of one year, which can be critical from the perspective of practical system for deployment. The novelty of this work lies in addressing effect of session variability and template aging to some extent with analysis. This knowledge can be utilized for a practical SV based framework under regular use for deployment.

The remaining paper is compiled in the following order: Section 2 explains the development of baseline SV system for the RedDots challenge. In Section 3 the proposed framework

http://dx.doi.org/10.21437/Interspeech.2016-1001

Table 1: *Baseline system performance on RedDots database*

Results of Male Subset								
Testing Condition	**Part I**		**Part II**		**Part III**		**Part IV TD**	
	EER (%)	**DCF**	**EER (%)**	**DCF**	**EER (%)**	**DCF**	**EER (%)**	**DCF**
Impostor True	9.50	0.0442	-	-	-	-	9.50	0.0433
Impostor Wrong	5.86	0.0268	6.58	0.0294	5.34	0.0242	6.09	0.0287
Results of Female Subset								
Impostor True	14.20	0.0524	-	-	-	-	13.73	0.0514
Impostor Wrong	9.78	0.0330	8.25	0.0353	3.79	0.0160	9.54	0.0355

for session variability study is explained which is found to exploit the session variation information for speaker modeling. Section 4 highlights how the template aging characteristics can be observed for data over a large session variation. Section 5 provides a discussion on the conducted study for session variability and template aging, which is pivotal for a deployable system with practical implementation. Finally, Section 6 provides the conclusion of the current work.

2. Development of Baseline Speaker Verification System for RedDots Challenge

This section describes the baseline system that is developed as a part of the challenge. The baseline framework is based on the i-vector modeling as front-end and probabilistic linear discriminant analysis (PLDA) as back-end module for SV architecture [13, 14].

2.1. Database

The RedDots database released for the challenge contains a population of 49 male and 13 female speakers totaling to 62 speakers having 572 sessions in total, that is collected with a collaborative effort of 21 countries across the world [12]. The database has four different parts out of which the Part I contains 10 fixed phrases that are common for all the speakers. The speakers of Part II and Part III have 10 unique phrases each of which are assigned and user chosen, respectively. While the Part IV of the database contains free text phrases that are unique across different sessions. For evaluation of Part IV there are two modalities, which are text-dependent (TD) and text-prompted (TP). Sessions 2, 4 and 6 are used for speaker modeling in each category of the dataset, except for TP case of the Part IV where sessions 1, 2, 3, 4 and 6 are used. This study is reported considering Part I to Part III and Part IV TD category of the RedDots database.

2.2. Pre-processing & feature extraction

The utterances from RedDots database are processed with blocks of 20 ms with a shift of 10 ms for every Hamming windowed frame and 39-dimensional (13 + 13Δ ++ 13$\Delta\Delta$) mel frequency cepstral coefficients (MFCC) are extracted. Energy based voice activity detection is performed to select the region of interest and then features of those regions are normalized with cepstral mean variance normalization (CMVN) technique [15].

2.3. i-vector and PLDA based framework for SV

The i-vector and PLDA based framework is used for the development of the baseline system for SV. The RSR2015 database is used as development data for building the universal background model (UBM), total variability matrix (T-matrix) and PLDA model [16]. Male and female data are processed separately and two gender dependent UBMs of 512 mixtures are built and then the statistics (zeroth and first order) are extracted for the development data as well as for the RedDots dataset. Two gender dependent T-matrix of 150-dimension are trained using development data statistics. The i-vectors of male and female subset are then extracted using respective T-matrix for the mentioned database. 100-dimensional PLDA model is trained using development data i-vectors for male and female subset separately, which is used for classification of a trial according to the evaluation procedure of RedDots Challenge.

Table 1 shows the performance of the baseline system developed for the RedDots challenge for all the four parts for male and female subsets. The results are reported in terms of equal error rate (ERR) and decision cost function (DCF) as per the evaluation protocol of RedDots database. In this work, the evaluation of trials are conducted under two conditions, impostors producing correct phrases (Impostor True) and impostors producing wrong phrases (Impostor Wrong). The testing condition based on genuine speakers producing wrong phrases (Target Wrong) is not considered as its scope is limited in a cooperative scenario for practical systems. The Part II and Part III of the RedDots database do not have the the Impostor True condition as these subsets concentrate of speaker-specific fixed phrases.

3. Proposed Framework for Session Variability Study

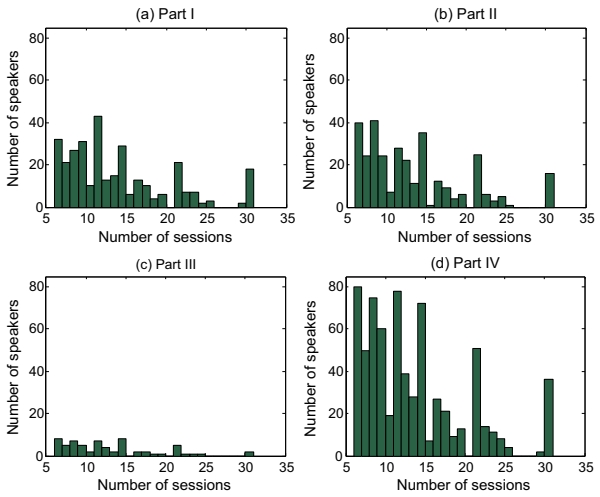

Figure 1: Histogram depicting number of sessions per speaker for male subset of RedDots database

The RedDots dataset is collected over a span of almost one year and hence there is a large session variability involved

Table 2: *Performance on RedDots database under implicit exploitation of session variability*

Results of Male Subset								
Testing Condition	**Part I**		**Part II**		**Part III**		**Part IV TD**	
	EER (%)	**DCF**	**EER (%)**	**DCF**	**EER (%)**	**DCF**	**EER (%)**	**DCF**
Impostor True	6.91	0.0324	-	-	-	-	6.67	0.0308
Impostor Wrong	4.41	0.0195	4.14	0.0186	3.40	0.0156	4.34	0.0199
Results of Female Subset								
Impostor True	9.15	0.0409	-	-	-	-	8.47	0.0387
Impostor Wrong	5.36	0.0213	4.30	0.0204	3.79	0.0214	4.72	0.0222

across the trials of each speaker. Figure 1 shows the histogram for number of sessions per speaker for the RedDots database for all the four parts. It is clearly visible that most of the speakers have more than 10-15 sessions of fixed phrase short utterances. However the data from Part III of the RedDots database has relatively less number of speakers having more number of sessions. This session variation in trials of the speakers has definitely produced an impact on the performance of the SV system, the results of which are shown in Table 1. The performance of Part I, Part II and Part IV are found to be poorer than Part III of the RedDots database as the trials are having more session variation for the former subsets.

The evaluation of RedDots database is made according to the evaluation procedure, where sessions 2, 4 and 6 from each speaker is used for modeling and remaining for testing. As the sessions 2, 4 and 6 are relatively former sessions compared to the sessions beyond 15-20 of each speaker, that may have produced a lot of mismatch in speaker characteristics as can be known from the literature review made in introduction section. To overcome this mismatch in session variation and to observe the effectiveness of it, an experimental setup is designed under which the first, middle and the last session of each speaker is taken for speaker modeling. As these sessions are used as test sessions in the baseline setup, the train sessions under baseline setup i.e. 2, 4 and 6 session of speakers are used for testing in the experimental setup which is currently designed for session variation study. The performance of the database for this experimental setup is then evaluated with i-vector PLDA based framework, the results of which can be seen in Table 2. The results show an significant improvement in performance than that of the baseline system in all parts of the RedDots database except very less for Part III due to the lesser number of sessions and speakers involved in Part III subset of the database. This improvement is mainly due to the implicit exploitation of session variability across the trials of each speaker, which is made by choosing the three sessions with large session variation for modeling the speakers. From the perspective of practical deployable system, where regular testing takes place from a population of speakers, this approach may be adopted over some sizeable interval to have an impact on SV performance by addressing session variability in an implicit manner.

4. Proposed Framework for Template Aging Study

The session variation in speakers is found to have an important aspect that reflects in system performance. As the RedDots database is collected over a period of almost one year, the effect of template aging is observed over it across all the speakers and this information may be useful from a practical system point of view, which can be studied. To study the effect of template aging of the speakers two experimental setups are designed that are explained in the following subsections.

4.1. First three sessions

In this study to observe the aging effect on RedDots database, the speaker models are generated considering the first three sessions of the speakers instead of sessions 2, 4 and 6 which are as per the baseline evaluation process. The trial list for testing under the baseline setup that contained the first three sessions are replaced with earlier train sessions, i.e. with 2, 4 and 6 session of the speakers. In this way the SV system is built over i-vector PLDA platform, the performance of which is mentioned in Table 3 for the male and female subsets. The results of this when compared to that of the baseline system performance, it can be seen that these are comparable to that obtained from the baseline, with slight improvement in almost all the cases.

4.2. Last three sessions

In the second set of studies for observing the effect of template aging, the last three sessions of the speakers are considered for speaker modeling and the system performance is evaluated in a similar manner as mentioned in the previous subsection for the study based on the first three sessions of the speakers. The results under this study using the last three sessions of speakers for modeling can be seen from Table 4. These results show that they are having significant improvement over the baseline system performance that is observed from Table 1 for the male and female subset of speakers. Additionally, the performance is far better from that obtained with training the speaker models with the first three sessions which is explained in previous subsection. Thus, the aging of speaker templates shows an interesting trend that can be useful for exploitation into a practical system.

5. Discussion

In this section, a discussion is made over the results obtained under session variability and template aging that is explained in Section 3 and Section 4 of this work. The studies under session variability showed that if the train templates of a speaker are taken from utterances of the speakers having larger session lapse among them, then that can help for achieving better performance. Similarly, the second set of study which is done for the aging of speaker templates shows that if the comparatively later sessions of a speaker are taken for modeling, that gives better performance than that obtained using earlier sessions with large gaps in time. Both these studies are beneficial from the perspective of deployable systems, as an update of speaker models may be done in the system. Further if we compare the performance of the study made with respect to session variation shown in Table 2 with the study of aging done with the last three sessions of speaker for modeling, which is showed in Table 4, the results are more or less comparable. This trend conveys that although there is not much session lapse in the last three sessions of a speaker, some vast speaker characteristics are evolved with aging of the speakers that are retained in

Table 3: *Performance on RedDots database considering first three sessions for modeling*

Testing Condition	Part I		Part II		Part III		Part IV TD	
	EER (%)	DCF	EER (%)	DCF	EER (%)	DCF	EER (%)	DCF
Results of Male Subset								
Impostor True	7.53	0.0376	-	-	-	-	7.65	0.0364
Impostor Wrong	4.53	0.0214	4.60	0.0215	4.53	0.0184	4.85	0.0234
Results of Female Subset								
Impostor True	13.56	0.0476	-	-	-	-	11.32	0.0447
Impostor Wrong	8.68	0.0299	5.33	0.0256	3.79	0.0214	7.40	0.0287

Table 4: *Performance on RedDots database considering last three sessions for modeling*

Testing Condition	Part I		Part II		Part III		Part IV TD	
	EER (%)	DCF	EER (%)	DCF	EER (%)	DCF	EER (%)	DCF
Results of Male Subset								
Impostor True	6.88	0.0326	-	-	-	-	6.57	0.0311
Impostor Wrong	4.53	0.0187	4.70	0.0216	3.88	0.0167	4.28	0.0198
Results of Female Subset								
Impostor True	10.25	0.0398	-	-	-	-	9.36	0.0376
Impostor Wrong	5.99	0.0238	4.98	0.0225	4.55	0.0203	5.70	0.0232

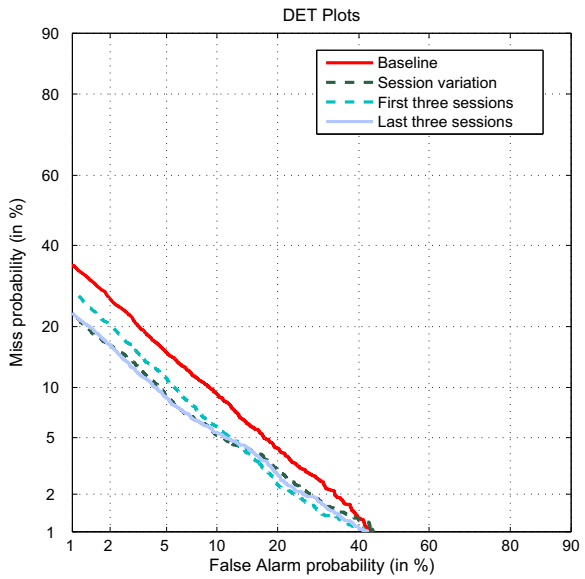

Figure 2: DET plot for different studies for Impostor True condition over male subset of Part I RedDots database

comparatively later sessions which have large gap from the earliest sessions. It has similar impact in the system performance if the speaker templates are chosen with large session variation among them for modeling. It is also to be noted that for Part I, Part II and Part IV of the database, the speakers are given fixed phrases that they have to produce unlike the Part III, where the speakers chose their fixed phrase themselves. This in turn effects in the SV performance as the speakers get acquainted with the phrases over time and struggle less to fix their pronunciation. If we assume that the speaker fixes his/her pronunciation after a few sessions, the three last sessions are closer from most of the sessions in term of pronunciation than the first three sessions. This makes the large session variation and template aging as interesting areas to explore together in coming years with more precise investigation into it.

Figure 2 shows the detection error tradeoff (DET) curves for different studies that have been made for session variability and template aging for Impostor True condition over Part I of male speaker subset of RedDots dataset. The plots clearly show that the baseline system performance is comparatively closer to that obtained from modeling the speakers with the first three sessions. While the performance that is obtained with modeling the speakers with first, middle and last session closely resembles to that obtained with modeling the speakers with the last three sessions, both the approaches being superior to the baseline performance. This highlights the matter of discussion in this section showing many interesting aspects of session variability and template aging that can be utilized for practical realizable systems.

6. Conclusion

This work makes an attempt to explore some issues related to session variability and template aging that can have significant impact from the outlook of a deployable system. The studies are carried over the RedDots database, which is collected for research on fixed phrase based short utterances having practical significance that includes data from about 21 countries that has sessions taken over a period of one year. A method for exploiting session variability in an implicit manner is proposed that outperforms the baseline system performance by a visible margin. Additionally, speaker template aging across sessions is investigated and an interesting trend is observed, that is imperative for a deployable system. The speaker models trained using the last three sessions of the speakers is proven to bear some advance speaker characteristics that evolved in a progressive manner with session variation. The studies made with respect to the effect of session variability and template aging can be utilized in real time SV systems to get the benefit out of it in regular sizeable intervals. In future we aspire to model session variability explicitly and exploit the aging effect across large session variation of speakers to observe the evolved vast characteristics in more detail.

7. Acknowledgement

This work is in part supported by a project grant 12(6)/2012-ESD for the project entitled “Development of Speech-Based Multi-level Person Authentication System” funded by the Department of Electronics and Information Technology (DeitY), Govt. of India.

8. References

[1] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification," *Computer Speech & Language*, vol. 22, no. 1, pp. 17 – 38, 2008.

[2] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," CRIM, Montreal, Tech. Rep. CRIM-06/08-13, 2005.

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

[4] 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.

[5] M. McLaren, R. Vogt, B. Baker, and S. Sridharan, "A comparison of session variability compensation approaches for speaker verification," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 4, pp. 802–809, Dec 2010.

[6] A. Stolcke, S. S. Kajarekar, L. Ferrer, and E. Shrinberg, "Speaker recognition with session variability normalization based on mllr adaptation transforms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 1987–1998, Sept 2007.

[7] A. Lanitis, "A survey of the effects of aging on biometric identity verification," *Int. J. Biometrics*, vol. 2, no. 1, pp. 34–52, Dec. 2010.

[8] H. Beigi, "Effects of time lapse on speaker recognition results," in *Digital Signal Processing, 2009 16th International Conference on*, July 2009, pp. 1–6.

[9] F. Kelly, A. Drygajlo, and N. Harte, "Speaker verification with long-term ageing data," in *Biometrics (ICB), 2012 5th IAPR International Conference on*, March 2012, pp. 478–483.

[10] J. P. Campbell, W. Shenand W. M. Campbelland R. Schwartzand J. F. Bonastre, and D. Matrouf, *Forensic speaker recognition*. IEEE Signal Processing Magazine, 2009, vol. 26, no. 2.

[11] Y. Matveev, *The Problem of Voice Template Aging in Speaker Recognition Systems*. Cham: Springer International Publishing, 2013, pp. 345–353.

[12] 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.

[13] 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.

[14] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *In Proc. Interspeech*, Aug 2011, pp. 249–252.

[15] 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.

[16] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, pp. 56 – 77, 2014.