Compensating Utterance Information in Fixed Phrase Speaker Verification

Rohan Kumar Das, Maulik Madhavi and Haizhou Li
Department of Electrical and Computer Engineering,
National University of Singapore, Singapore
E-mail: {rohankd, elemaul, haizhou.li}@nus.edu.sg

***Abstract*—This work investigates on explicitly utilizing utterance information for fixed phrase speaker verification (SV). In this scenario, the same phrase is spoken by the speakers during the training and testing sessions. In other words, the speaker model possesses both speaker as well as utterance information. Therefore, there is a potential to improve the speaker characterization by compensating the utterance information. In this work, we propose a framework to compensate the utterance information, which is used to normalize the lexical content. A hidden Markov model (HMM) based triphone model is considered as a universal background model (UBM). It is used for adapting the speaker-utterance model and the background utterance model in the proposed utterance compensation framework. Given a test utterance and a claimed speaker-utterance model, the UBM as well as background utterance model is utilized for compensating the average speaker information and the lexical content information, respectively. The studies are conducted on RSR2015 database, which reveal the importance of the proposed utterance compensation framework as compared to the framework without utterance compensation.**

I. Introduction

Speaker verification (SV) refers to verifying a speaker's claimed identity with reference to a given speech. It can be broadly classified into text-dependent and text-independent category [1], [2]. The former deals with production of same lexical contents by the speakers during training and testing sessions. Typically, fixed phrases of 2-3 seconds are considered for this kind of framework. On the contrary, there is no restriction on lexical content to be produced in text-independent SV, which therefore requires a larger amount of speech data for modeling and verification of a trial to achieve a benchmark performance. In case of real-world scenario, computational time involved for speaker authentication is a key factor. Thus, text-dependent SV emerges as a suitable candidate for application oriented systems.

The text-dependent SV has gained attention in the community in recent years with the availability of standard large databases like RSR2015 and RedDots [3], [4]. Earlier work in this field focuses on dynamic time warping (DTW) based temporal alignment technique that utilizes the sequence information [5]. However, with the recent developments, different advanced techniques are carried out in this domain. For instance, a hierarchical multi-layer acoustic model (HiLAM) is proposed based on Gaussian mixture model (GMM)-hidden Markov model (HMM) architecture in [3]. This model captures speaker as well as lexical sequence information and is found to perform better than the i-vector based speaker modeling that dominates in text-independent SV [6]. Some of the other novel frameworks proposed in this direction include joint factor analysis (JFA) [7], i-vector/HMM [8] and unsupervised HMM-universal background model (UBM) [9] which have produced effective performance. Further, the use of deep learning models, including deep neural network (DNN) with restricted Boltzmann machine (RBM) and DNN/i-vector have been studied for text-dependent SV in [10] and [11], respectively.

As discussed, the text-dependent SV advocates on the usage of same fixed phrase for verification of a trial. Thus, the lexical content has a definite significance for this kind of SV. The phonetic posteriorgram feature and model based frameworks can utilize the lexical information for speaker modeling as found in the literature [12], [13]. Along similar directions, the speaker and lexical content have been modeled jointly in terms of a HMM triphone model followed by usage of different alignment strategies for an improved text-dependent SV system in [14]. The authors of [15] have performed utterance verification and SV separately followed by their combination at decision and score level that improves SV performance. Additionally, studies have shown that it is effective to normalize the content information in case of text-dependent SV. The work in [16] depicts that mismatch in content severely degrades performance and posterior normalization helps to deal with the mismatch. Further, the authors of [17] represent the speakers in terms of phone classes for i-vectors obtained for each senone unit. In [18], a content normalization strategy is applied on the extraction of posterior probabilities that improves performance. These studies show that modeling lexical content with speaker modeling as well as lexical content normalization, both helps in text-dependent SV system. This motivated for exploring these two directions in a common framework for fixed phrase SV.

In this work, the speaker-utterance models are created using a HMM based triphone model to jointly capture the speaker and utterance information as given in [14]. Further, in order to compensate the lexical content information we introduce a background utterance model. We then propose an utterance compensation framework using the background utterance model to normalize the lexical information from speaker-utterance model for an improved fixed phrase SV system. This kind of framework is suitable for real-world

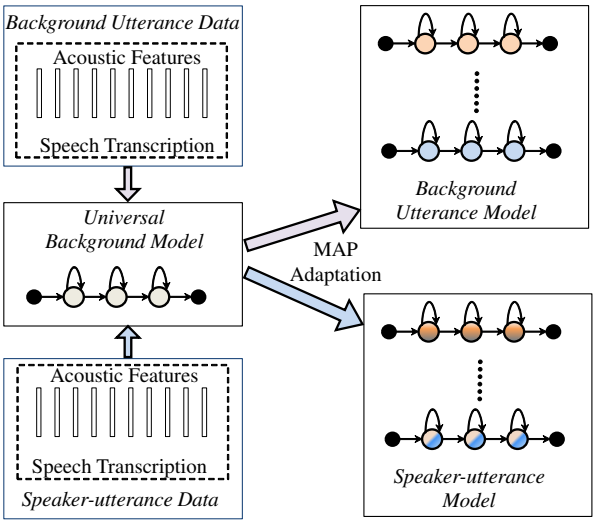

Fig. 1. Training phase for the proposed utterance compensation framework.

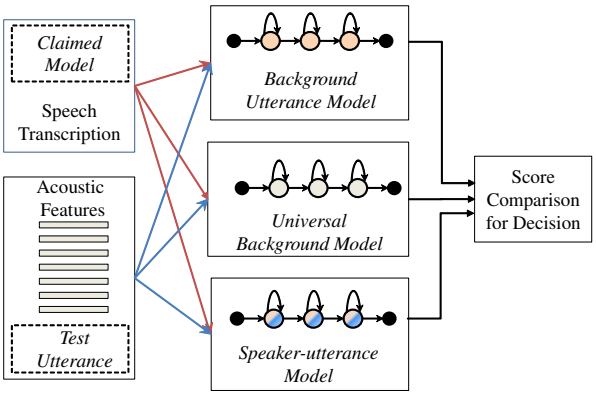

Fig. 2. Testing phase for the proposed utterance compensation framework.

scenario, where different speakers have to enroll in a system using multiple fixed phrases and during testing one of the phrases is prompted for authentication. The RSR2015 database having multiple fixed phrases and conditions is considered for the studies. The novelty of this work is attributed in proposing the framework of utterance compensation for text-dependent SV systems.

The reminder of the paper is organized as follows. Section II describes the architecture of the proposed utterance compensation framework. In Section III, the details regarding the development of the fixed phrase SV system is mentioned. Section IV reports the results of the work carried out along with a discussion. Finally, the paper in concluded in Section V.

II. Utterance Information Compensation

This section describes the proposed utterance compensation framework. In this framework, to jointly capture the speaker and utterance information, we first obtain the speaker-utterance model as suggested in [14]. Then a background utterance model is introduced to normalize the lexical content information from the speaker-utterance model. Next we discuss the training and testing phase of the proposed framework.

A. Training Phase

The lexical information in a speech signal can have different levels of realization, such as phones, syllable or words. A monophone model doesn't take the phonetic context into consideration during acoustic modeling. However, the triphone model exploits the contextual phonetic information and hence, it is more suitable for acoustic modeling [19]. In this regard, a triphone based HMM is used as a UBM to capture the lexical variations from a wide range of speakers. The adaptation is performed with respect to this background model using a maximum a posteriori (MAP) approach. The MAP adaptation utilizes the use of prior knowledge about the acoustic model. With the help of the prior informative knowledge and background acoustic model, MAP adaptation can be used to estimate the parameters from target acoustics data. In absence of the prior information, the estimates of MAP are identical to the maximum-likelihood (ML) approach [20]. The objective of MAP adaptation is to maximize the posterior, that is achieved in two steps, namely, computation of sufficient statistics and adapt to the old statistics [21]. The first step computes the sufficient statistics for the background data feature vector sequence $\mathcal{O}^{bkg} := \{\mathbf{o}_1, \mathbf{o}_2, \dots \mathbf{o}_T\}$,

$$\tilde{\mu}_{jm} = \frac{\sum_{t=1}^{T} \gamma_{jm}(t)\mathbf{o}_t}{\sum_{t=1}^{T} \gamma_{jm}(t)} \tag{1}$$

where, j is the HMM state, m the mixture component index, T is the number of feature vectors, and $\gamma_{jm}(t)$ is the sufficient statistics weight. In other words, $\gamma_{jm}(t)$ is the posterior probability of data vector $\mathbf{o}_t$ being in j^{th} HMM state and m^{th} mixture component. The second step adapts the old statistics with the relevance factor τ. In a practical scenario, the adaptation of mean component is mostly considered in MAP adaptation.

In this work, we have used MAP adaptation on UBM to derive the speaker-utterance model and background utterance model. The mean parameters after MAP adaptation for speaker-utterance model having feature vector sequence $\mathcal{O}^{su} := \{\mathbf{o}_1^{su}, \mathbf{o}_2^{su}, \dots \mathbf{o}_{T_{su}}^{su}\}$ is expressed as:

$$\mu_{jm}^{su} = \frac{\tau\tilde{\mu}_{jm} + \sum_{t=1}^{T_{su}} \gamma_{jm}(t)\mathbf{o}_t^{su}}{\tau + \sum_{t=1}^{T_{su}} \gamma_{jm}(t)} \tag{2}$$

where, T_{su} is the total number of feature vectors used for the speaker-utterance model adaptation and $\tilde{\mu}_{jm}$ is the old statistics for mean vector obtained from the background dataset as computed by Equation (1) that needs to be adapted. Similarly, the mean parameters after MAP adaptation for background utterance model having feature vector sequence $\mathcal{O}^{utt} := \{\mathbf{o}_1^{utt}, \mathbf{o}_2^{utt}, \dots \mathbf{o}_{T_{utt}}^{utt}\}$ is expressed as:

$$\mu_{jm}^{utt} = \frac{\tau\tilde{\mu}_{jm} + \sum_{t=1}^{T_{utt}} \gamma_{jm}(t)\mathbf{o}_t^{utt}}{\tau + \sum_{t=1}^{T_{utt}} \gamma_{jm}(t)} \tag{3}$$

where, T_{utt} is the total number of feature vectors used for the background utterance model adaptation.

Figure 1 shows the block diagram of training phase for the proposed utterance compensation framework. It shows that the UBM is adapted to obtain speaker-utterance model using the data (i.e., transcription and acoustic features) for a speaker-utterance pair. Similarly, the background data is used to create the background utterance model via MAP adaptation approach as explained in this section. The background utterance model is expected to capture the gross lexical information of the particular phrase from a set of background speakers.

B. Testing Phase

During testing, the test feature vector sequence $\mathcal{O}^{te} := \{\mathbf{o}_1^{te}, \mathbf{o}_2^{te}, \dots \mathbf{o}_{T_{te}}^{te}\}$ and speech transcript of the claimed model $\mathcal{W}$ are used to compute log-likelihood scores with respect to the three different acoustic HMMs. These are the claimed speaker-utterance model (λ_{su}), UBM (λ_{ubm}) and background utterance model (λ_{utt}) that are trained as described in Section II-A. The scores obtained from each of these models are used in combination for verification of a trial.

In consideration of background utterance model in the proposed utterance compensation framework, the scoring criteria for test data $\mathcal{O}^{te}$ against the claimed speaker-utterance model having word transcription $\mathcal{W}$ is computed as follows:

$$\mathcal{S}_{\mathcal{O}^{te}}^{\mathcal{W}} = \log P(\mathcal{O}^{te}|\lambda_{su}, \mathcal{W}) - \frac{1}{2}\Big[\log P(\mathcal{O}^{te}|\lambda_{ubm}, \mathcal{W}) + \log P(\mathcal{O}^{te}|\lambda_{utt}, \mathcal{W})\Big] \quad (4)$$

The likelihood scores are computed with respect to each model by using the forward-backward algorithm to sum up the likelihood values for each HMM state. For instance, the log-likelihood with reference to the UBM for testing data $\mathcal{O}^{te}$ can be computed as follows:

$$\log P(\mathcal{O}^{te}|\lambda_{ubm}, \mathcal{W}) = \frac{1}{T_s}\sum_{t\in\mathcal{T}} \log \sum_j P(\mathbf{o}_t^{te}|q_t^{ubm} = j) P(\mathbf{o}_t^{te}|\theta^j, q_t^{ubm} = j) \quad (5)$$

where, $P(\mathbf{o}_t^{te}|q_t^{ubm} = j)$ represents the state alignment probability for feature vector $\mathbf{o}_t^{te}$ with respect to the j^{th} HMM state of the UBM and $P(\mathbf{o}_t^{te}|\theta^j, q_t^{ubm} = j)$ represents the likelihood computed using background parameters θ^j. Further, $\mathcal{T}$ indicates the set of frames (time) for which the HMM-state sequence aligns to the non-silence phone and T_s is the total number of frames belonging to non-silence phones. The silence does not carry any information related to either speaker or lexical content and rather the likelihood values associated with the silence frames create confusion in SV. Similarly, the log-likelihood against speaker-utterance model (i.e., $\log P(\mathcal{O}^{te}|\lambda_{su}, \mathcal{W})$) and background utterance model (i.e., $\log P(\mathcal{O}^{te}|\lambda_{utt}, \mathcal{W})$) can be computed to obtain the final score for decision making as given by Equation (4). In the absence of the background utterance model, the proposed framework resembles to the joint speaker-utterance framework.

Figure 2 shows the block diagram of testing phase for the proposed utterance compensation framework. The speech transcription of the claimed speaker-utterance model and acoustic feature vectors of the test utterance are fed to three different models to compute respective likelihood scores as shown in Figure 2. The scores obtained for speaker-utterance model is compared against the scores from UBM and the background utterance model. To balance the compensation effect between the UBM and the background utterance model, we used equal contribution from them by averaging their scores. The details regarding the parameters and experimental setup are described in the next section.

III. System Description

In this section, the details of the SV system developed in this work are mentioned. The database, front-end processing and experimental setup are described in the following subsections.

A. Database

The RSR2015 database [3] is used for conducting the studies in this work. It contains a population of 300 speakers involving 157 male and 143 female speakers. Based on the type of phrases used, the database is categorized in three parts, namely, Part I, Part II and Part III. The Part I subset consists of 30 different fixed phrases that are of duration around 3-4 seconds. The 30 fixed short command based phrases are kept in Part II that ranges in 1-2 seconds in duration. Finally, the 13 random digit sequence based phrases are grouped under Part III. Each phrase is spoken for 9 different sessions by all the speakers, out of which sessions 1st, 4th and 7th are used for training. The remaining sessions are used for evaluating the performance.

Further, the database is divided in three different sets namely, background, development and evaluation sets. The background set is used for developing background models, whereas the development and evaluation sets are used to evaluate the performance. There are 50 male and 47 female speakers in the background set, 50 male and 47 female speakers in development set and the remaining 57 male and 49 female speakers are included in evaluation set. Additionally, the database has three different trial categories, namely, *Impostor Correct*, *Target Wrong* and *Impostor Wrong*. However, this work focuses on *Impostor Correct* category, where the impostors utter the correct phrase for validating a claim. This has been chosen as this condition is suitable in a practical cooperative scenario and represents the reply attack scenario as well.

B. Front-end Processing

The short term processing is performed on speech utterances from the RSR2015 database with frame size of 20 ms and a shift of 10 ms. 60-dimensional (20-base + 20-Δ + 20-$\Delta\Delta$) mel frequency cepstral coefficient (MFCC) features including energy coefficient are extracted for every frame considering 23 logarithmically placed mel filters. There is no voice activity detection applied as the SV framework used in this work

TABLE I
PERFORMANCE IN TERMS OF EER (%) FOR DIFFERENT FRAMEWORKS ON PART I OF RSR2015 DATABASE.

System	Development Set		Evaluation Set	
	Female	Male	Female	Male
HiLAM [3]	3.24	3.69	2.96	2.47
Joint Speaker-utterance	1.60	2.05	1.01	1.39
Utterance Compensation	1.64	**1.46**	**0.73**	**0.96**
Utterance Compensation with u^f	**1.46**	**1.46**	**0.72**	**0.96**

TABLE II
PERFORMANCE IN TERMS OF EER (%) FOR DIFFERENT FRAMEWORKS ON PART II OF RSR2015 DATABASE.

System	Development Set		Evaluation Set	
	Female	Male	Female	Male
HiLAM [3]	6.66	10.58	7.95	8.38
Joint Speaker-utterance	4.10	4.92	3.24	4.26
Utterance Compensation	**4.03**	**4.16**	**2.85**	**3.61**
Utterance Compensation with u^f	**3.86**	**4.16**	**2.79**	**3.61**

considers information from non-silence frames after HMM state alignment as mentioned in Section II-B. The features of each utterance are normalized in the cepstral domain with cepstral mean and variance normalization (CMVN) [5].

C. Experimental Setup

As discussed in Section II, the UBM is a triphone based HMM. This has been trained using the entire background set of RSR2015 database except the first session based utterances from different speakers for each phrase. These first session based utterances are kept aside for the creation of background utterance model. The rationale behind this is to avoid overlap of data considered for utterance model creation with UBM. Additionally, we would like to highlight that the background utterance model is created from the background data of RSR2015 database, which does not have any overlap of speakers with development and evaluation set. Each phone is modeled as a left-to-right HMM (Bakis model) with three emitting states and total number of Gaussian components is kept as 512. In this work, we used CMU pronunciation dictionary converting speech transcription into sequence of phones. The phone set comprises of 39 phones and 1 silence phone. The number of tied states (or senones) obtained after decision tree clustering is 429. We have followed the standard WSJ recipe of Kaldi to create the UBM, which is a triphone based HMM [22].

It is to be noted that gender dependent modeling is performed to have separate systems for male and female speakers. The trained UBM is then used to create speaker-utterance model and background utterance model for Part I and Part II of RSR2015 database as described in Section II-A. Given a test trial and a claimed speaker-utterance model, the scoring is performed as mentioned in Section II-B. The evaluation of the system performance on RSR2015 database follows the standard protocol depicted in [3].

IV. RESULTS AND DISCUSSION

This section reports the results carried out with respect to the current work along with a discussion. The HiLAM system described in [3] is considered as a common reference system for comparison and results are cited from the same. In addition, we consider the joint speaker-utterance system that considers only speaker-utterance model and UBM as suggested in [14]. This framework is different from the HiLAM [3] in terms of HMM based modeling. In joint speaker-utterance system, the speaker-utterance model a triphone based HMM obtained with transcription, whereas the HiLAM does not use any transcription. In addition, HMM parameters are adapted from GMM-UBM in the case of HiLAM. The proposed utterance compensation framework introduces a background utterance model to normalize the lexical information from joint speaker-utterance system. Table I shows the performance for the stated frameworks on Part I of RSR2015 database in terms of equal error rate (EER). It can be observed that the joint speaker-utterance framework outperforms the conventional HiLAM. The possible reason behind this can be that the triphone based HMM used in joint speaker-utterance system captures speaker-specific information in a better way than the HMM involved in HiLAM. Additionally, the utterance compensation framework further enhances the performance, except one case for female development set, showing its significance for fixed phrase SV.

The studies are extended to carry out on Part II of RSR2015 database that deals with short commands of 1-2 seconds. Table II shows the performance for different frameworks on Part II of RSR2015 database considered in this work. The results under this database subset also show the effectiveness of the utterance compensation framework. Further, it is to be noted that the difference in performance between HiLAM and the proposed framework is more in case of Part II subset having short commands. This reflects that the importance of the utterance information compensation is even more for fixed phrases of very short duration.

It is observed from Table I and Table II that the contribution of the proposed utterance compensation framework is relatively less for the results obtained on the development set for female speakers in case of Part I and Part II subsets. As discussed in Section II-B, equal weightage has been applied to UBM and background utterance model during final score computation for decision making. However, we argue that there may be a better strategy to have weighted compensation of these models. In this regard, we introduce an utterance factor (u^f), that is learned on the development set and then it is applied on the evaluation set. The Equation (4) is modified to include the utterance factor as follows:

$$\begin{aligned}\mathcal{S}^{\mathcal{W}}_{\mathcal{O}^{te}} &= \log P(\mathcal{O}^{te}|\lambda_{su},\mathcal{W}) \\ &\quad - \left[(1-u^f)\log P(\mathcal{O}^{te}|\lambda_{ubm},\mathcal{W}) + u^f \log P(\mathcal{O}^{te}|\lambda_{utt},\mathcal{W})\right]\end{aligned} \tag{6}$$

where, the utterance factor u^f is a scalar that ranges between 0 and 1. Its optimal value u^f_{opt} is obtained in steps of 0.1 to have the least cost.

The experiments on development set are performed to compute the u^f for both male and female subsets of RSR2015 database and then applied on the evaluation set. An utterance

factor of $u^f_{opt} = 0.5$ is obtained for the male set, whereas for female speakers $u^f_{opt} = 0.3$. Thus, it shows that although for male speakers the equal compensation strategy is optimal, for female speakers the weighted combination with utterance factor is found to be optimal for minimum cost. The last row of Table I and Table II represents the results associated with utterance factor based hypothesis to have improved results for female speakers. The proposed utterance compensation framework is beneficial for practical systems that has multiple fixed phrases for enrollment and randomly one out of them is used for verification of a claim. In this study, we used GMM/HMM framework to derive state posterior probability. The same can be extended to DNN for deriving the posterior probability as used in [13] for utterance compensation framework. Additionally, the future work will focus on extending the studies presented in this work for prompted digit sequence based SV.

V. Conclusions

We propose a novel framework to compensate the lexical content from speaker-utterance models in a fixed phrase based SV. The lexical content is found to carry critical information for text-dependent SV. In order to utilize this information, firstly a speaker-utterance model is created via adaptation from a HMM based triphone model. We then hypothesize that compensating the utterance information can improve the SV performance. A background utterance model is created for every fixed phrase from the background data to support the hypothesis. During testing, the test utterance is compared to the claimed speaker-utterance model along with the UBM as well as background utterance model. The background utterance model is used to normalize the average utterance information for improved speaker characterization in the proposed utterance compensation framework. The studies are conducted on RSR2015 database that portray the importance of the utterance compensation for fixed phrase SV. Furthermore, it is observed that the proposed framework is more effective towards short commands compared to the existing systems in the domain of text-dependent SV.

VI. Acknowledgements

This research 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] M. Hèbert, "Text-dependent speaker recognition," *Springer-Verlag Heidelberg*, pp. 743–762, 2008.

[2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, pp. 12 – 40, 2010.

[3] 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.

[4] 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.

[5] 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.

[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] P. Kenny, T. Stafylakis, J. Alam, P. Ouellet, and M. Kockmann, "Joint factor analysis for text-dependent speaker verification," in *Odyssey 2014, Joensuu, Finland*, 2014, pp. 200–207.

[8] H. Zeinali, H. Sameti, L. Burget, J. ernock, N. Maghsoodi, and P. Matjka, "i-vector/HMM based text-dependent speaker verification system for reddots challenge," in *Interspeech 2016, San Francisco*, 2016, pp. 440–444.

[9] A. K. Sarkar and Z.-H. Tan, "Text dependent speaker verification using un-supervised HMM-UBM and temporal GMM-UBM," in *Interspeech 2016, San Francisco*, 2016, pp. 425–429.

[10] Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang, and K. Yu, "Deep feature for text-dependent speaker verification," *Speech Communication*, vol. 73, pp. 1–13, 2015.

[11] J. Yan, L. Xie, G. Wang, and Z. Fu, "A segmental DNN/i-vector approach for digit-prompted speaker verification," in *Asia-Pacific Signal and Information Processing Association APSIPA ASC 2017, Kuala Lumpur, Malaysia*, 2017, pp. 1–5.

[12] 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.

[13] S. Dey, S. Madikeri, M. Ferras, and P. Motlicek, "Deep neural network based posteriors for text-dependent speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2016, Shanghai, China*, March 2016, pp. 5050–5054.

[14] G. Wang, K. A. Lee, T. H. Nguyen, H. Sun, and B. Ma, "Joint speaker and lexical modeling for short-term characterization of speaker," in *Interspeech 2016, San Francisco*, 2016, pp. 415–419.

[15] T. Kinnunen, M. Sahidullah, I. Kukanov, H. Delgado, M. Todisco, A. K. Sarkar, N. B. Thomsen, V. Hautamki, N. Evans, and Z.-H. Tan, "Utterance verification for text-dependent speaker recognition: A comparative assessment using the reddots corpus," in *Interspeech 2016, San Francisco*, 2016, pp. 430–434.

[16] N. Scheffer and Y. Lei, "Content matching for short duration speaker recognition," in *Interspeech 2014, Singapore*, 2014, pp. 1317–1321.

[17] L. Chen, K. A. Lee, B. Ma, W. Guo, H. Li, and L. Dai, "Phone-centric local variability vector for text-constrained speaker verification," in *Interspeech 2015, Dresden, Germany*, 2015, pp. 229–233.

[18] S. Dey, S. Madikeri, P. Motlicek, and M. Ferras, "Content normalization for text-dependent speaker verification," in *Interspeech 2017, Stockholm, Sweden*, 2017, pp. 1482–1486.

[19] S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*. Cambridge University Press, 2006.

[20] A. R. Webb, *Statistical pattern recognition*. John Wiley & Sons, 2003.

[21] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding 2011*, Dec. 2011.