INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Content Normalization for Text-dependent Speaker Verification

Subhadeep Dey[1,2], *Srikanth Madikeri*[1], *Petr Motlicek*[1] and *Marc Ferras*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
{subhadeep.dey, srikanth.madikeri, petr.motlicek,marc.ferras}@idiap.ch

Abstract

Subspace based techniques, such as i-vector and Joint Factor Analysis (JFA) have shown to provide state-of-the-art performance for fixed phrase based text-dependent speaker verification. However, the error rates of such systems on the random digit task of RSR dataset are higher than that of Gaussian Mixture Model-Universal Background Model (GMM-UBM). In this paper, we aim at improving i-vector system by normalizing the content of the enrollment data to match the test data. We estimate i-vectors for each frames of a speech utterance (also called online i-vectors). The largest similarity scores across frames between enrollment and test are taken using these online i-vectors to obtain speaker verification scores. Experiments on Part3 of RSR corpora show that the proposed approach achieves 12% relative improvement in equal error rate over a GMM-UBM based baseline system.

Index Terms: speaker verification, i-vectors, content matching

1. Introduction

The state-of-the-art techniques in Speaker Verification (SV) such as i-vector and Joint Factor Analysis (JFA) have shown to provide high performance for a variety of conditions including long duration utterances [1, 2]. When applied to forensics or voice-based access control, systems are often asked to deal with short recordings of speech. However, the performance of text-independent SV systems on short test utterances is far from being acceptable for any deployable system [3]. The performance can be enhanced considerably by constraining the speakers to utter a specific phrase [4, 5]. This form of authentication is referred to as text-dependent SV.

There are various strategies to implement a text-dependent system. In fixed-phrase based text-dependent SV, the phrase of the test data is expected to be identical to the enrollment (as shown in column 1 of Table 1). In case it is not, the system can reliably detect the mismatch and reject the claim. In many text-dependent applications, we would like to impose lesser constraint on the speaker while maintaining the same level of accuracy of the fixed-phrase based systems. In one of the scenarios, the words of the test phrase are subset of the content of the enrollment. A potential example is when speaker models are created by pooling all N phrases uttered by the speaker during enrollment, while during test phase, the speaker utters only one of the N phrases. Experiments in [6] show that the state-of-the-art i-vector system performs worse for this task compared to the fixed phrase based SV.

In this paper, we are interested in designing a SV system to better understand the effect of content in these two text-dependent scenarios:
(a) *Seen*: We create the scenario as considered in [6] by using the phrases from RSR dataset. The enrollment data is created by pooling all the phrases spoken by the speaker. The test data consists of a single phrase as illustrated in Table 1 (Column 2), and
(b) *Random-Digits*: the enrollment phase consists of the speaker uttering permutations of ten digits. During testing, the speaker is prompted to utter five digits only as shown in Table 1 (Column 3).

Table 1: *A valid enrollment-test phrase pair for text-dependent speaker verification systems for different tasks. We use sample phrases from RSR dataset.*

Tasks	Enrollment phrase	Test phrase
Fixed-phrase	"the redcoats ran like rabbits"	"the redcoats ran like rabbits"
Seen	{ "the redcoats ran like rabbits", "only lawyers love millionaires", $\cdots$ }	any of the enrollment phrases
Random-Digits	{ "five", "four", $\cdots$, "ten" }	{ "two", "five", $\cdots$ }

Various techniques have been explored that aim at exploiting the content information of the test data for *Seen* and *Random-Digits* tasks [7, 8, 6]. In [7], content information is used by extracting an i-vector for every linguistic unit of the utterance for the *Random-Digits* task. It has been shown that significant gain in performance can be achieved using this approach. In [6], posteriors estimated using a Deep Neural Network (DNN) are used for i-vector extraction for the *Seen* task. This approach outperforms a Gaussian Mixture Model (GMM) based i-vector system, as the DNN is trained for content discrimination. Furthermore, an approach that scales sufficient statistics of the enrollment to match test statistics is proposed as a way to successfully deal with content mismatch [6].

The approaches described above perform content matching in the i-vector framework using context-dependent state (senone) posteriors estimated using DNN. Nevertheless, estimating senone posteriors from Automatic Speech Recognition (ASR) word recognition lattices instead of the DNN forward pass improves the performance of the i-vector system for text-independent SV system [9]. These senone posteriors incorporate the information of both the acoustic (incorporating also lexical model) and language models. In this work, we apply the senone posteriors estimated from ASR word recognition lattices for the *Seen* and *Random-Digits* tasks.

In the past, selecting common set of words or phones between the enrollment and test utterance [10, 11] have shown to increase SV performance. We refer to the process of transforming the enrollment utterance to match the lexical content as content normalization. We present an approach to perform content normalization by selecting regions explicitly in the enrollment data to match the test data by employing speaker in-

http://dx.doi.org/10.21437/Interspeech.2017-1419

formative features. In our previous work [12], we found that features estimated using i-vector extractor (also termed as online i-vectors) are beneficial for the fixed phrase task. We use the online i-vectors for the *Seen* and *Random-Digits* tasks as it has been shown to contain speaker-content informative characteristics [12].

The paper is organized as follows: Section 2 presents the baseline systems while Section 3 describes SV using posteriors generated by ASR and the content normalization technique. Sections 4 and 5 describe the experimental setup for the evaluating the system and discuss the achieved results by various systems. Finally, the paper is concluded in Section 6.

2. Baseline Systems

The state-of-the-art text-independent SV approach to model speakers is built around total variability subspace technique [2]. This approach assumes that the invariant speaker characteristics lie in a low dimensional subspace of mean GMM supervectors. A speaker model is represented by a fixed-dimensional vector called *i-vector*.

In [6], DNNs were used to cluster the acoustic space into linguistic units such as senones, making it easier to focus on the content of each utterance. The posterior probabilities of each of the senones were then used for i-vector extraction. A posterior normalization technique was further proposed to scale the zero-th and first order statistics of the enrollment data to match those in the test data [6]. The technique is described as follows. Let N_e and N_t be the zero-th order statistics of the enrollment and test utterances respectively, and $\mathbf{F}_e$ and $\mathbf{F}_t$ be the first order statistics of the enrollment and test utterances respectively. The new statistics for the enrollment are obtained as

$$\mathrm{N}_e^{'} = \beta \mathrm{N}_e \tag{1}$$

$$\mathbf{F}_e^{'} = \beta \mathbf{F}_e, \tag{2}$$

where β is a normalization constant, which is defined as $\mathrm{N}_t/\mathrm{N}_e$. When N_e or N_t is 0, β is set to zero as well. The details of the technique can be found in [6]. We consider the following as the baseline systems, (a) GMM-Universal Background Model (UBM), and (b) i-vector system using the posterior normalization technique.

3. Posteriors and Content Matching

In this work, we use two techniques to perform content normalization, (a) one based on DNN posterior estimation and (b) using online i-vectors. Both are described in the following section.

3.1. Posteriors from ASR decoder

An i-vector system involves the estimation of zero-th and first order statistics as a prior step to computing the i-vectors. The state-of-the-art SV systems compute these statistics using the senone posteriors obtained at the output of the DNN [6, 13]. Therefore, the DNN acts as a short-term content estimator in terms of senones.

In this work, senone posteriors are obtained after decoding using language and lexical models, in the context of an ASR system. In [9], it was shown that senone posteriors obtained after ASR decoding performed better than those obtained after a DNN forward pass. The former posteriors are smoothed by using language constraints and drastically improve the phone accuracy.

In our work, we use a lattice decoder [14], based on a Weighted Finite State Transducer (WFST), that outputs a graph of hypothesized sequences of words. Senone posterior probabilities are estimated from the acoustic scores at the nodes of the lattice, after the forward-backward recursion, for each frame. These are used for i-vector extraction. For content normalization, we use the posterior normalization technique as proposed for the baseline system [6].

3.2. Content normalization using i-vectors

In the past, strategies to exploit phonetic information have been successful for text-dependent SV. In [7], i-vectors are extracted for each of the senone units, which are then clustered to obtain speaker representation for the phone classes for *Random-Digits* task. In [6], they analyze the performance of i-vector system for *Seen* task. Experiments using state-of-the-art techniques show that content mismatch has a strong impact on the SV performance [6] and normalizing posteriors reduces the error rate considerably. Recent results show that selecting common linguistic units between enrollment and test data produces low error rate [11, 15] for text-independent SV. Motivated by these results, we hypothesize that normalizing the content of the enrollment data with speaker and content informative features will be beneficial for the *Seen* and *Random-Digits* tasks.

In our previous work [12, 16], we used online i-vectors as features to Dynamic Time Warping (DTW) algorithm for fixed phrase based text-dependent SV task. Significant gain in performance was observed as opposed to using the conventional i-vectors which suggests that these features contain sufficient speaker and content information. We use online i-vectors as features for performing content normalization.

The strategy to perform content normalization is as follows. Online i-vectors are estimated for each speech frame with a context of 10 frames (i.e. sufficient statistics are estimated with a window size of 21 frames). This leads to a sequence of online i-vectors corresponding to an utterance. Enrollment and test content are matched by computing the maximum similarity scores from each online i-vector in test to all instances in enrollment. As many scores as the number of speech frames in test utterance are obtained. Finally, these scores are averaged to obtain a global similarity score. The rationale behind this approach is to choose the closest frame in the enrollment data. The accumulated global score is obtained as follows

$$s(\mathbf{X},\mathbf{Y}) = \frac{1}{C}\sum_j min\{d(\mathbf{x}_i,\mathbf{y}_j), \forall i = \{1, 2, \cdots, R\}\}, \tag{3}$$

where $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_R\}$ and $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_C\}$ represent set of i-vectors for the enrollment and test data, the function $\mathrm{d}(\mathbf{x}_i, \mathbf{y}_j)$ computes the distance between the i-vectors $\mathbf{x}_i$ and $\mathbf{y}_j$. The score $s(\mathbf{X},\mathbf{Y})$ represents the accumulated distance between the closest speech frames. We used cosine distance metric to compute the dissimilarity between two online i-vectors. A threshold on the cosine distance can be applied to detect if a test frame is not present in the enrollment data.

The content normalization technique described above does not assume phonetic label of the speech frame. In a scenario, when phonetic alignments are obtained using the text-transcripts, the minimization of Equation 3 could be performed by iterating over the same phonetic category of the enrollment data.

3.3. PLDA as a feature extractor

The online i-vector representation contains other information in addition to the speaker content. In order to factor out the channel effects, a PLDA model is trained as the back-end classifier with online i-vectors as features. In our previous work [12], PLDA trained with speaker-phone pairs was used for fixed phrase based text-dependent SV task. In this paper, we explore speaker-word combination as classes definition for the training the PLDA. A speech recognizer is employed to align the development data with the word labels. Online i-vectors corresponding to word boundaries are subsequently used as features for the PLDA model. The PLDA model is then used to project the online i-vectors using the parameters of the model to obtain channel compensated vectors as done in [17, 12]. We refer to these vectors as plda-vectors.

4. Experimental Setup

In this section, we describe the experimental setup for the baseline and proposed systems.

4.1. Evaluation and Training Data

We performed experiments on Part1 and Part3 portion of the RSR dataset [18, 5, 19], restricting to female speakers only. We evaluated our systems on these two text-dependent tasks:
(a) *Seen*: We created the following test set as described in [6] to evaluate our techniques. The data of each of the speakers involves 15 pass-phrases with three sessions for each pass-phrase, for a total of 45 utterances. The total duration of the enrollment of a speaker is 90 s. The test utterance consists of a speaker uttering a phrase with a duration of 2 s. For this task, the evaluation trials consist of 4’410 target and 211’680 impostor trials. For both the tasks, the Fisher female subset English was used as the training data. It contains about 1.3 k utterances with 120 hours of speech data. For the *Seen* task, the Speaker Recognition Evaluation (SRE) data from SRE 04 to 08 was used for training the back-end classifier.
(b) *Random-Digits*: This subset contains 49 speakers pronouncing random sequence of digits. The protocol described in [18] was adopted to perform text-dependent SV. Three utterances (with an average duration of 12 s) are used for creating the enrollment model. The enrollment utterance consists of the speaker uttering 10 digits. The test utterance consists of 5 digits with an average duration of 2 s. For this task, the evaluation trials consists of 5’283 target and 253’584 impostor trials. The Part3 of RSR dev portion was used as the development data. We used 1’264 utterance consisting of 47 speakers pronouncing ten digits.

4.2. I-vector system

The front-end SV system extracts Mel Frequency Cepstral Coefficients (MFCC) of 20 dimensions from 25 ms of frame of speech signal with 10 ms sliding window and delta, double delta features appended to it. Short time gaussianization is applied to the features using a 3 s sliding window [20, 21]. The dimensionality of i-vector extractor is set to 400.

4.3. ASR system

DNN acoustic model is trained as a part of the ASR system. It is trained with MFCCs with 4 hidden layers each of dimension 1’200. The output layer has 1.9 k senone units including 20 silence units. The same ASR system is designed for both tasks.

Table 2: *Performance of the different baseline systems in terms of EER (%). The* ***GMM-UBM*** *provides the best performance among the baseline systems in both evaluation tasks.*

Systems/Tasks	*Seen*(%)	*Random-Digits* (%)
Ivec$^{\mathbf{GMM}}_{\mathbf{PLDA}}$	16.5	17.3
Ivec$^{\mathbf{DNN}}_{\mathbf{PLDA}}$	11.6	15.2
PN-Ivec$^{\mathbf{GMM}}_{\mathbf{PLDA}}$	12.3	15.8
PN-Ivec$^{\mathbf{DNN}}_{\mathbf{PLDA}}$	8.6	14.4
GMM-UBM	**4.5**	**8.6**

It employs a CMU dictionary with 42 k words, similar to [3]. The ASR system is validated on a separate subset consisting of 200 utterances from the Fisher database with 3gram word LM. The Word Error Rate (WER) on the validation set is 24.4%. The senone posteriors extracted from the DNN forward pass are used to estimate the parameters of the i-vector model. We used the conventional ASR decoder parameters to obtain word recognition lattices [14] (beam width of 13). The same type of lattices has been used previously for various tasks [22, 23, 24]. From these lattices, we obtain the senone posteriors, by fixing the acoustic scale parameter to 0.01, in order to obtain i-vectors that follow a Gaussian distribution. Furthermore, we observed that higher acoustic scale (> 0.01) leads to i-vectors with high kurtosis and thus making the PLDA model ineffective.

5. Experimental Results and Discussions

In this section, we describe the results obtained with the baseline and the proposed SV systems. The various systems considered in this paper are the following:

- **GMM-UBM**: a universal GMM is created using the training data (UBM). The speaker models are obtained from this UBM using Maximum-a-Posteriori (MAP) adaptation.
- **Ivec**$_{\mathbf{PLDA}}$: the conventional i-vector systems for speaker recognition. The systems using GMM, DNN and decoded ASR lattice posteriors are referred to as **Ivec**$^{\mathbf{GMM}}_{\mathbf{PLDA}}$, **Ivec**$^{\mathbf{DNN}}_{\mathbf{PLDA}}$ and **Ivec**$^{\mathbf{DNN\text{-}dec}}_{\mathbf{PLDA}}$ respectively.
- **PN-Ivec**$_{\mathbf{PLDA}}$: the systems using posterior normalization technique as explained in Section 3.1. The systems using GMM, DNN and decoded ASR lattice posteriors for i-vector extraction are referred to as **PN-Ivec**$^{\mathbf{GMM}}_{\mathbf{PLDA}}$, **PN-Ivec**$^{\mathbf{DNN}}_{\mathbf{PLDA}}$ and **PN-Ivec**$^{\mathbf{DNN\text{-}dec}}_{\mathbf{PLDA}}$ respectively.
- **CN-Ivec**: the SV systems applying content normalization technique using i-vectors as explained in Section 3.2. The systems using GMM, DNN and decoded ASR lattice posteriors for i-vector extraction are referred to as **CN-Ivec**$^{\mathbf{GMM}}$, **CN-Ivec**$^{\mathbf{DNN}}$ and **CN-Ivec**$^{\mathbf{DNN\text{-}dec}}$ respectively.
- **CN-Ivec**$^{\mathbf{DNN}}_{\mathbf{PLDA}}$: a PLDA model is trained on top of the online i-vectors as the channel compensation model. We explore the use of speaker-phone and speaker-word pairs to train the PLDA. The systems trained on plda-vectors (estimated using online i-vectors with DNN and decoded ASR posteriors) with speaker-phone pairs are referred to as **CN-Ivec**$^{\mathbf{DNN}}_{\mathbf{PLDA,p}}$ and **CN-Ivec**$^{\mathbf{DNN\text{-}dec}}_{\mathbf{PLDA,p}}$, while the systems trained on plda-vectors trained with speaker-word labels are referred to as **CN-Ivec**$^{\mathbf{DNN}}_{\mathbf{PLDA,w}}$ and **CN-Ivec**$^{\mathbf{DNN\text{-}dec}}_{\mathbf{PLDA,w}}$

Table 3: *Performance of the different SV systems (using senone posteriors extracted from decoded ASR lattices) in terms of EER(%). The* $\textbf{PN-Ivec}^{\text{DNN-dec}}_{\text{PLDA}}$ *performs the best among the other systems for Seen task.*

Systems/Tasks	*Seen (%)*	*Random-Digits (%)*
$\textbf{Ivec}^{\text{DNN-dec}}_{\text{PLDA}}$	10.9	18.9
$\textbf{PN-Ivec}^{\text{DNN-dec}}_{\text{PLDA}}$	**5.6**	15.7

5.1. Baseline SV systems

Table 2 shows the performance of various i-vector and GMM-UBM based SV systems for the *Seen* and *Random-Digits* tasks. We observe that performance of the systems on *Seen* is significantly worse than the fixed phrase based text-dependent system [12]. Lower bound for *Seen* task is 2.3% Equal Error Rate (EER) for the case when the phrases of the enrollment are identical to the test [12]. The posterior normalization technique is used to exploit the content of the enrollment data. We observe that this approach reduces the error rates by 26% relative (11.6% to 8.6% absolute) and 5% relative (15.2% to 14.4% absolute) EER for the *Seen* and *Random-Digits* tasks. Furthermore, we observe that incorporating the phonetic information (with DNN and decoded ASR posteriors) helps the SV. The **GMM-UBM** provides the best performance among the baseline systems considered in this paper. The EER for this system is comparable to the results published in literature [25, 7]. We applied T-norm on the scores produced by the **GMM-UBM** system. We observe that T-norm reduces from 10.5% to 8.6% absolute EER for the *Random-Digits* task.

5.2. SV systems using ASR lattice posteriors

We explore the application of posteriors estimated from word recognition ASR lattices in an i-vector framework. Table 3 shows the performance of the i-vector systems using these posteriors. We observe that $\textbf{Ivec}^{\text{DNN-dec}}_{\text{PLDA}}$ outperforms $\textbf{Ivec}^{\text{DNN}}_{\text{PLDA}}$ for *Seen* task by 0.7% absolute EER. Significant gain in performance is achieved by the $\textbf{PN-Ivec}^{\text{DNN-dec}}_{\text{PLDA}}$ compared to $\textbf{PN-Ivec}^{\text{DNN}}_{\text{PLDA}}$, with 35% relative (8.6% to 5.6% absolute) EER for *Seen*. This indicates the importance of more accurate senone alignments in obtaining better SV performance for this task. However, performance of $\textbf{Ivec}^{\text{DNN-dec}}_{\text{PLDA}}$ and $\textbf{PN-Ivec}^{\text{DNN-dec}}_{\text{PLDA}}$ degrade for the *Random-Digits* task compared to the $\textbf{Ivec}^{\text{DNN}}_{\text{PLDA}}$. One of the reasons could be that the performance of the ASR system (unconstrained LM) is poor on the RSR dataset ($\sim$ 80% WER).

5.3. SV systems based on content normalization technique

As opposed to using posterior normalization, we also explore content normalization using i-vectors, as described in Section 3.2. Table 4 shows the performance of the proposed content normalization based SV systems using posteriors from GMM, DNN and decoded ASR lattices. We observe that the proposed systems outperform the posterior normalization based systems in *Seen* and *Random-Digits* tasks. In particular, the $\textbf{CN-Ivec}^{\text{DNN}}$ performs better than $\textbf{PN-Ivec}^{\text{DNN}}_{\text{PLDA}}$ by 67% relative (8.6% to 2.8% absolute) and 15% relative (14.4% to 12.2% absolute) EER for the *Seen* and *Random-Digits* tasks respectively. This indicates the importance of the content normalization technique using online i-vectors. We observe that $\textbf{CN-Ivec}^{\text{DNN}}_{\text{PLDA,p}}$ performs better than the **GMM-UBM** by 10% relative (8.6% to 7.7% absolute) EER. The $\textbf{CN-Ivec}^{\text{DNN}}_{\text{PLDA,w}}$ further improves upon $\textbf{CN-Ivec}^{\text{DNN}}_{\text{PLDA,p}}$ by 0.2% absolute EER in *Random-Digits* task. Thus, training the PLDA using speaker-word labels is more effective in the random digits task than the speaker-phone pairs. We do not present all the results of content normalization technique using plda-vectors with GMM, DNN and decoded ASR posteriors as we did not obtain better performance than $\textbf{CN-Ivec}^{\text{DNN}}_{\text{PLDA,w}}$.

Table 4: *Performance of the different SV systems (using content normalization technique) in terms of EER(%). The* $\textbf{CN-Ivec}^{\text{DNN}}_{\text{PLDA,w}}$ *performs the best among the other systems in Seen task. The * indicates the system using text-transcript.*

Systems/Tasks	*Seen (%)*	*Random-Digits (%)*
$\textbf{CN-Ivec}^{\text{GMM}}$	4.1	13.4
$\textbf{CN-Ivec}^{\text{DNN}}$	2.8	12.2
$\textbf{CN-Ivec}^{\text{DNN-dec}}$	4.3	15.5
$\textbf{CN-Ivec}^{\text{DNN}}_{\text{PLDA,p}}$	**2.7**	7.7
$\textbf{CN-Ivec}^{\text{DNN}}_{\text{PLDA,w}}$	**2.7**	**7.5**
$\textbf{CN*-Ivec}^{\text{DNN}}_{\text{PLDA,w}}$	**2.5**	7.6

We also explore the importance of the text-transcript for the content normalization technique. An ASR system is used to align the enrollment and test data with the ground truth. Scores from the closest frames between the enrollment and test data are accumulated by iterating over same phonetic classes. The EER for the *Seen* task reduces by 0.2% absolute for the $\textbf{CN-Ivec}^{\text{DNN}}_{\text{PLDA,w}}$ system. However, for the *Random-Digits* task, we did not get any improvement compared to **7.5**% EER.

6. Conclusions

In this paper, we address a text-dependent SV task in which the lexical content of the test data has been spoken by the speaker. The conventional approach to tackle this problem is to incorporate content information in the i-vector framework using senone posteriors (estimated from DNN). A posterior normalization technique is applied to scale the sufficient statistics of the enrollment data to match the statistics of the test data. Significant gain in performance is observed for the *Seen* task compared to the baseline i-vector system.

We proposed to improve upon the baseline system by, (a) enhancing the senone prediction accuracy of the DNN posteriors, and (b) normalizing the content of the enrollment to match the test using online i-vectors. We explore the use of speaker-word pair to train the PLDA model on top of online i-vectors. The PLDA is used to obtain channel compensated vectors (plda-vectors). We observe that content normalization using plda-vectors achieves the best results for *Seen* and *Random-Digits* tasks with 40% and 12% relative EER over a baseline **GMM-UBM** system.

7. Acknowledgements

This work was supported by the EU FP7 project Speaker Identification Integrated Project (SIIP).

8. References

[1] D. G. Romero and C. Y. E. Wilson, “Analysis of ivector length normalization in speaker recognition systems,” in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27 to 31, 2011*, 2011, pp. 249–252.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *Audio,*

Speech, and Language Processing, IEEE Transactions on, vol. 19, no. 4, pp. 788–798, May 2011.

[3] P. Motlicek *et al.*, "Employment of subspace gaussian mixture models in speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4445–4449.

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

[5] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Modelling the alternative hypothesis for text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 734–738.

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

[7] L. Chen, K. A. Lee, B. Ma, W. Guo, H. Li, and L.-R. Dai, "Phone-centric local variability vector for text-constrained speaker verification," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] H. Aronowitz and O. Barkan, "On leveraging conversational data for building a text dependent speaker verification system." in *INTERSPEECH*, 2013, pp. 2470–2473.

[9] H. Su and S. Wegmann, "Factor analysis based speaker verification using asr," *Interspeech 2016*, pp. 2223–2227, 2016.

[10] M. Hébert, "Text-dependent speaker recognition," in *Springer handbook of speech processing*. Springer, 2008, pp. 743–762.

[11] A. Stolcke, E. Shriberg, L. Ferrer, S. Kajarekar, K. Sonmez, and G. Tur, "Speech recognition as feature extraction for speaker recognition," in *Signal Processing Applications for Public Security and Forensics, 2007. SAFE'07. IEEE Workshop on*. IEEE, 2007, pp. 1–5.

[12] S. Dey, P. Motlicek, S. Madikeri, and M. Ferras, "Template-matching for text-dependent speaker verification," *Speech Communication*, vol. 88, no. C, pp. 96–105, 2017.

[13] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 1695–1699.

[14] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, and G. Stemmer, "The kaldi speech recognition toolkit," in *In IEEE 2011 workshop*, 2011.

[15] B. J. Baker, R. J. Vogt, and S. Sridharan, "Gaussian mixture modelling of broad phonetic and syllabic events for text-independent speaker verification," 2005.

[16] S. Dey, P. Motlicek, S. Madikeri, and M. Ferras, "Exploiting sequence information for text-dependent speaker verification," in *ICASSP 2017*. IEEE, March 2017.

[17] S. Dey, S. Madikeri, and P. Motlicek, "Information theoretic clustering for unsupervised domain adaptation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, March 2016.

[18] A. Larcher, K. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, pp. 56–77, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2014.03.001

[19] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Phonetically-constrained plda modeling for text-dependent speaker verification with multiple short utterances," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7673–7677.

[20] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification." In Proc. of Speaker Odyssey, 2001, pp. 213–218.

[21] S. Madikeri, S. Dey, M. Ferras, P. Motlicek, and I. Himawan, "Idiap submission to the nist sre 2016 speaker recognition evaluation," Idiap, Tech. Rep., 2016.

[22] P. Motlicek, F. Valente, and I. Szoke, "Improving acoustic based keyword spotting using lvcsr lattices," in *Proceedings on IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2012, pp. 4413–4416.

[23] P. Motlicek, P. N. Garner, N. Kim, and J. Cho, "Accent adaptation using subspace gaussian mixture models," in *The 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2013.

[24] D. Imseng, P. Motlicek, P. N. Garner, and H. Bourlard, "Impact of deep mlp architecture on different acoustic modeling techniques for under-resourced speech recognition," in *Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding*, 2013.

[25] T. Stafylakis, P. Kenny, J. Alam, and M. Kockmann, "Jfa for speaker recognition with random digit strings."