EXPLOITING SEQUENCE INFORMATION FOR TEXT-DEPENDENT SPEAKER VERIFICATION

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

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

ABSTRACT

Model-based approaches to Speaker Verification (SV), such as Joint Factor Analysis (JFA), i-vector and relevance Maximum-a-Posteriori (MAP), have shown to provide state-of-the-art performance for text-dependent systems with fixed phrases. The performance of i-vector and JFA models has been further enhanced by estimating posteriors from Deep Neural Network (DNN) instead of Gaussian Mixture Model (GMM). While both DNNs and GMMs aim at incorporating phonetic information of the phrase with these posteriors, model-based SV approaches ignore the sequence information of the phonetic units of the phrase. In this paper, we tackle this issue by applying dynamic time warping using speaker-informative features. We propose to use i-vectors computed from short segments of each speech utterance, also called online i-vectors, as feature vectors. The proposed approach is evaluated on the RedDots database and provides an improvement of 75% relative equal error rate over the best model-based SV baseline system in a content-mismatch condition.

Index Terms— Text-dependent speaker verification, DNN posteriors, Dynamic Time Warping

1. INTRODUCTION

In the past few years, the state-of-the-art Speaker Verification (SV) systems have shown to provide high performance for long duration speech recordings [1, 2]. In practical applications (forensics, biometrics, etc.), SV is often applied on short duration test utterances ($\sim$3 s). However, results of the SV systems on short duration test set are yet to reach acceptable range of performance of any deployable system [3]. Unlike unconstrained scenarios, application of SV systems on constrained content of the test utterances can bring reasonable performance. Real applications have usually used phrases, digits and short commands to constrain the content [4, 5]. In this paper, we focus on text-dependent SV with fixed phrases being shared across speakers. The SV system is expected to verify the combination of claimed speaker and phrase. The trials, where either the speaker or phrase does not match, are considered as impostors. In this case, impostors can be divided into three categories, (i) the content does not match (content-mismatch), (ii) the speaker does not match (speaker-mismatch), and (iii) neither the speaker nor the content matches.

Most of the techniques to tackle text-dependent SV can be grouped into two main categories: (a) model-based and (b) template-based (Dynamic Time Warping (DTW)) techniques. The model-based SV techniques have mostly leveraged from text-independent SV solutions [6, 4]. They involve mainly the subspace-based formulation to the SV problem. Unlike these approaches, DTW attempts to match the enrollment and test templates of feature vectors.

More particularly, several works have explored relevance Maximum-a-Posteriori (MAP), i-vector or Joint Factor Analysis (JFA) for text-dependent SV [6, 7]. Relevance MAP adaptation of Gaussian Mixture Models (GMM) assumes that the speaker can be represented as a shift of the means and variances of a Universal Background Model (UBM) [8]. The i-vector technique has been used with a back-end classifier, Probabilistic Linear Discriminant Analysis (PLDA), trained on speaker-phrase combinations [4, 7]. In the same line, JFA using speaker-phrase and session terms has been shown to perform well [6]. The major improvements in model-based approaches for the fixed-phrase based text-dependent SV systems are achieved by incorporating Deep Neural Network (DNN) in the i-vector and JFA models [7, 9]. DNN is employed to estimate posteriors of the phonetic units, as a replacement for the GMM-UBM. Considering that a phrase can be decomposed into phonetic units and their sequences, DNNs provide a way to incorporate phonetic information of the phrase in model-based SV systems. However, information of the sequence of phonetic units in the phrase is still ignored. We believe that exploiting this sequence information will enhance system performance as the text-constraints are being used in the process.

In the past, various techniques have been proposed to exploit the sequence information of the phrase [4, 7]. In [4], a universal Hidden Markov Model (HMM) was used for adapting to each of the enrollment phrases. In another direction, DTW algorithm was proposed to match sequences of input features [7]. Besides the conventional spectral feature vectors, DTW can be applied on posterior vectors estimated using GMMs or DNNs. In [7, 10], it has been shown that such systems can outperform model-based SV systems (relevance MAP, i-vector, JFA) in content-mismatch condition. Although in speaker-mismatch condition, the DTW systems did not reach performance of model-based SV systems [7, 10], we presume this was due to applying features that are not necessarily speaker-discriminative. In this paper, we propose to extract speaker informative features using i-vector model, to be subsequently used as input to the sequence-matching algorithm. In particular, we extract i-vectors on short segments of each speech utterance ($\sim$200 ms), also referred to as "online i-vectors".

The online i-vectors have recently been successfully used as features for Automatic Speech Recognition (ASR) and speaker diarization task [11, 12]. In this paper, we propose to use online i-vectors as features to the DTW algorithm, especially to improve speaker-mismatch SV condition.

The paper is organized as follows: Sections 2 and 3 describe the baseline system and the proposed DTW approach, respectively. Sec-

tion 4 describes the experimental setup for evaluating the SV systems and Section 5 presents the results of the various proposed systems. Finally, conclusions are drawn in Section 6.

2. BASELINE SYSTEMS

The conventional systems for text-dependent SV use GMM based speaker modeling [6, 7, 4], i.e., it assumes that the data of a speaker is generated by a GMM. The following baseline systems are considered in this paper, (i) relevance MAP, (ii) i-vector and (iii) JFA. These are referred to as model-based SV approaches in this work.

2.1. Relevance MAP

In relevance MAP framework, the parameters of a GMM-UBM are estimated by pooling the data from many speakers [8]. To enroll a new speaker, the parameters of the GMM-UBM are adapted to match the speaker data under the MAP criterion. In practice, adapting only the means has shown to be sufficient [8]. To verify a claim against a speaker, the likelihood of the utterance is computed with respect to the model and compared against the likelihood with respect to the GMM-UBM.

2.2. I-vector

In the i-vector framework [2], the mean supervector of the GMM-UBM adapted to an utterance is transformed using a low-rank total variability matrix, as given by the following equation

$$\mathbf{s} = \boldsymbol{\mu} + \mathbf{T}\mathbf{w}, \tag{1}$$

where $\mathbf{s}$ is the speaker mean supervector and $\boldsymbol{\mu}$ is the mean supervector of a GMM-UBM. The matrix $\mathbf{T}$ defines a low-rank projection of the mean supervectors. The posterior mean of $\mathbf{w}$ is a low-dimensional vector called i-vector. A PLDA model is usually trained on top of i-vectors, exploiting multiple speaker-phrase combinations as input labels [4].

2.3. JFA

JFA is an alternative to the i-vector approach for speaker modeling. JFA can be built to explicitly model and compensate for the content variability as a separate factor [13, 14]. The JFA model is given as follows

$$\mathbf{s} = \boldsymbol{\mu} + \mathbf{D}\mathbf{z} + \mathbf{U}\mathbf{x}\,, \tag{2}$$

where $\mathbf{D}$ is a diagonal matrix capturing the speaker variabilities, $\mathbf{z}$ is the corresponding latent vector representing the speaker-phrase variability, $\mathbf{U}$ is the Eigenchannel matrix and $\mathbf{x}$ is the corresponding channel factor representing the channel effects of a speech recording.

2.4. DNN based SV systems

The posterior probabilities computed while estimating an i-vector or JFA factors assume feature vectors to be generated by a GMM. It has been shown that extracting posteriors from linguistically meaningful units, as opposed to the traditional GMM, can significantly improve speaker recognition performance [9]. A DNN acoustic model estimates the context-dependent tied-state (also called senones) posteriors, to be subsequently used for i-vector extraction in the SV task. Incorporating DNN posteriors in the i-vector and JFA frameworks have been found useful for text-dependent SV as well [7].

3. DYNAMIC TIME WARPING ALGORITHM

Section 2 briefly described the relevance MAP, i-vector and JFA approaches for text-dependent SV with fixed phrases. The phonetic information related to the phrase is captured by posteriors estimated using GMMs and DNNs. However, the sequence information of phonetic units of the phrase is ignored. We hypothesize that exploiting the sequence information of the phonetic units in addition to the content information can be helpful in achieving better performance as the text-constraints of the task are being used in the SV process.

In [4], the sequence information of the phrase is captured by using a universal HMM built from the training data. The speaker-phrase dependent model is obtained by MAP adaptation of the universal HMM. In our past work, we have already shown that a non-parametric method such as DTW can be used efficiently to capture sequence information [7]. The DTW is a dynamic programming technique to compute the distance between two sequences, and is commonly used in spoken word detection and other data-mining tasks [15, 16]. DTW approaches to text-dependent SV have generally used spectral vectors [17]. Posteriors from GMMs and DNNs have been recently used successfully for this task [10, 7]. In particular, senone posteriors obtained at the output of the DNN have shown to perform significantly better than the i-vector or JFA system in content-mismatched trials. However, in speaker-mismatched condition, the i-vector and JFA systems performed better [7]. The main reason was due to the input features for DTW were not necessarily speaker discriminative. DTW is supposed to align enrollment and test speech while the features are assumed to capture the speaker characteristics. Hence this paper proposes to extract speaker informative features (i.e. online i-vectors derived from short segments of speech data) for DTW, expecting an improved performance on the SV task.

Online i-vectors have been successfully used in speaker diarization and ASR systems [12, 11]. In ASR, online i-vectors have been applied as an additional input to the acoustic model training and adaptation by appending them with conventional spectral vectors [11]. In speaker diarization, these features have been concatenated with spectral features, which were subsequently fed to the speaker clustering algorithm [12]. In both works, a reasonable gain in performance was achieved compared to using only spectral features. This suggests that online i-vector representation contains sufficient speaker information. In the view of these results, we propose to use online i-vectors as an input to the DTW algorithm. The following section describes the process of extracting online i-vectors.

3.1. Online i-vectors

Figure 1 illustrates the process of extracting online i-vectors from the speech signal. Let the speech utterance contain 'M' frames of speech given by $\mathbf{O} = \{\mathbf{o}_1, \cdots, \mathbf{o}_\mathrm{M}\}$, where $\mathbf{o}_t$ is the t^{th} speech frame. The online i-vector corresponding to t^{th} speech frame of an utterance is computed with a context size of L frames. The Sufficient Statistics (SS) required for i-vector estimation are computed from the sequence of speech frames, starting from t - L to t + L. For a context size $\mathrm{L} = 10$ frames, a sliding window of $2L + 1$ frames is used with a shift step of 1 frame. Windows are centered at each frame of the utterance, which results in fewer frames being considered at the utterance boundaries. The corresponding sequence of online i-vectors is represented by $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_\mathrm{M}\}$ for an utterance. To compare two sequences of online i-vectors and obtain a similarity score, the cosine distance metric is applied in DTW algorithm.

The DTW score computed using online i-vectors is expected to reflect both content and speaker similarities between enrollment and test templates. A window length of 200 ms, corresponding to average syllable duration, is able to capture both types of information.

4. EXPERIMENTAL SETUP

In this section, we describe the experimental setup for the baseline and proposed systems, as well as system configurations of the

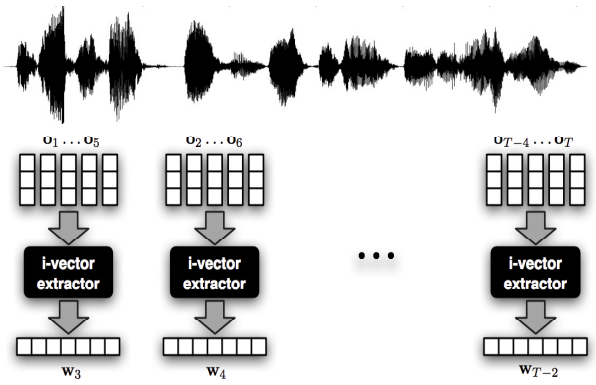

Fig. 1: Extraction of online i-vectors.

i-vector, JFA and DNN acoustic model. Evaluations are done on three conditions labeled as Cond1, Cond2, Cond3, and an additional condition (Cond-all) with the trials from all three conditions put together. More particularly, in condition 1, each trial is associated with determining if the phrases are the same or different. In condition 2, the system is required to differentiate speakers pronouncing the same content. In condition 3, both the speaker and the phrase can be different. Performance is presented in terms of Equal Error Rate (EER) and minimum Decision Cost Function (minDCF) with the probability of target being 0.01, cost of false alarm error probability being 1 and cost of miss error probability being 10, according to the protocol in [4]. In this paper, we evaluate our systems on the male part of the RedDots dataset. The training and development data are also taken from the male part of different databases.

4.1. Training, Development and Evaluation data

All experiments in this paper are performed on 8 kHz speech files. The training data is drawn from the subset of Fisher corpus ($\sim$120 h). It contains 1.2K utterances with an average duration of 5 minutes per utterance. We choose the Part1 of RSR [4] as the development data, which contains 42'305 utterances from 157 speakers. We evaluated our systems on the Part1 portion of the RedDots database [5] (downsampled to 8 kHz). This dataset contains 52 sessions per speaker, with one session recorded per week. This database is challenging in terms of long-term intra-speaker variability compensation in addition to inter-speaker variability. It consists of 35 speakers with speech from 10 fixed pass-phrases, which are different from the phrases of the RSR dataset. The evaluation contains a total of 3'242 target trials and 1'230'038 impostor trials, out of which conditions 1, 2 and 3 contain 29'178, 120'086 and 1'080'774 impostor trials, respectively.

4.2. I-vector and JFA systems configurations

Twenty dimensional Mel Frequency Cepstral Coefficients (MFCC) are extracted from 25 ms of speech frames with 10 ms sliding window, appended with delta and acceleration parameters. Short time gaussianization is applied to these features using a 3 s sliding window [18]. The Hungarian phoneme recogniser is used to detect voice activity. It compares the sum of posteriors over all phone classes with the posterior of the silence class to classify each frame as speech or non-speech [19]. This is used to mark the start and end points of the speech region in the audio. The parameters of the 1'024-component GMM-UBM are estimated using the training dataset mentioned in Section 4.1. The i-vector extractor of 400 dimensions is also trained with the same training set. The parameters of the JFA system are estimated on the development dataset using speaker-phrase labels. The rank of the eigenchannel matrix $\mathbf{U}$ (of Equation 2) is set to 50.

Table 1: Performance of the various GMM based baseline systems on RedDots dataset in terms of EER(%). The $\mathbf{RMAP}^{\mathbf{GMM}}$ outperforms other baseline systems across all conditions.

Systems/Conditions	Cond1	Cond2	Cond3	Cond-all
$\mathbf{RMAP}^{\mathbf{GMM}}$	**5.2**	**4.1**	**1.0**	**1.8**
$\mathbf{Ivec}^{\mathbf{GMM}}_{\mathbf{PLDA}}$	6.9	4.2	1.3	1.9
$\mathbf{JFA}^{\mathbf{GMM}}$	10.5	7.9	2.9	3.8

Table 2: Performance of the various DNN-based systems on RedDots dataset in terms of EER(%).

Systems/Conditions	Cond1	Cond2	Cond3	Cond-all
$\mathbf{Ivec}^{\mathbf{DNN}}_{\mathbf{PLDA}}$	6.9	**3.4**	**1.2**	**1.6**
$\mathbf{JFA}^{\mathbf{DNN}}$	**4.1**	7.0	**1.2**	2.5

4.3. DNN acoustic model

The DNN acoustic model used in this paper to estimate senone posterior probabilities is trained in an ASR fashion with the data as described in Section 4.1. Alignments for training are obtained from a HMM/GMM ASR system developed on the same data. We used the Kaldi toolkit to train a DNN with 4 hidden layers with 1'200 sigmoid units per layer and 1'530 softmax units at the output. The input to the DNN are 660 dimensional vectors obtained by stacking 11 MFCC feature vectors. The accuracy of the HMM/DNN acoustic model is validated on a speech recognition task using a separate Fisher subset consisting of 200 utterances. The ASR system employs a CMU dictionary with 42k words and a 3-gram language model, similar to [3]. We achieved Word Error Rate (WER) of about 31%. This DNN is subsequently used to compute the posteriors of the senone units as a prior step to estimating the parameters of i-vector and JFA models.

5. RESULTS

In this section, we describe the results on a SV task obtained for the baseline and the proposed systems on RedDots database. Since the relevance MAP approach has shown to provide good results on RedDots data [20], we consider this system as the baseline. First, we analyse the performance of the model-based SV approaches and then the DTW systems.

The following SV systems will be analysed:

- $\mathbf{RMAP}^{\mathbf{GMM}}$: the speaker models are obtained from GMM-UBM by MAP adaptation.
- $\mathbf{Ivec}_{\mathbf{PLDA}}$: the conventional i-vector-PLDA SV systems developed using GMM-UBM or DNN SS, which are referred to as $\mathbf{Ivec}^{\mathbf{GMM}}_{\mathbf{PLDA}}$ and $\mathbf{Ivec}^{\mathbf{DNN}}_{\mathbf{PLDA}}$ respectively.
- **JFA**: this system represents Joint Factor Analysis model. The JFA systems exploiting GMM-UBM and DNN SS are referred to as $\mathbf{JFA}^{\mathbf{GMM}}$ and $\mathbf{JFA}^{\mathbf{DNN}}$ respectively.
- **DTW**: sequences of speech feature vectors (MFCCs) and senone posterior vectors estimated using the GMM-UBM or DNN are compared using the DTW algorithm. The DTW systems with MFCCs, GMM posteriors and DNN posteriors are referred to as **DTW-MFCC**, $\mathbf{DTW\text{-}post}^{\mathbf{GMM}}$ and $\mathbf{DTW\text{-}post}^{\mathbf{DNN}}$ respectively.

Table 3: Performance of the various DTW systems on RedDots dataset in terms of EER(%).

Systems/Conditions	Cond1	Cond2	Cond3	Cond-all
DTW-MFCC	2.1	5.6	1.2	1.9
DTW-post^{GMM}	1.8	6.7	1.7	2.9
DTW-post^{DNN}	**1.1**	9.0	1.0	3.5
DTW-onIvec^{GMM}	2.6	3.8	1.3	1.8
DTW-onIvec^{DNN}	1.3	**3.2**	**0.8**	**1.3**
onIvec$^{GMM}_{PLDA}$	5.4	6.7	2.5	2.9
onIvec$^{DNN}_{PLDA}$	2.8	4.8	1.8	2.1

- **DTW-onIvec**: this system employs sequences of online i-vectors as an input to DTW algorithm. The online i-vectors are estimated using SS either from GMM-UBM or DNN, which are referred to as **DTW-onIvec^{GMM}** and **DTW-onIvec^{DNN}** respectively.

5.1. Model-based SV systems

Table 1 compares performance of relevance MAP, i-vector and JFA systems exploiting posteriors estimated using GMM-UBM. It can be seen that **RMAP^{GMM}** achieves the best results among the model-based SV systems, which is consistent with the results in [20]. Although in [14], the **JFA^{GMM}** outperformed **RMAP^{GMM}** and **Ivec$^{GMM}_{PLDA}$** on the RSR database, we obtain contradictory performance on RedDots data. One of the reasons maybe that the JFA model (trained on speaker-phrases of RSR database) overfits to the training phrases of RSR.

As explained in Section 2, **Ivec$^{DNN}_{PLDA}$** and **JFA^{DNN}** benefit from incorporating a linguistic information incorporated by DNN. The DNN acoustic model estimates the senone posteriors used in i-vector extraction process. The top 10 scoring posteriors are retained for building i-vector extractor and JFA. It can be observed from Table 2 that incorporating DNN posteriors in the i-vector and JFA systems consistently improves the SV performance. For instance, **Ivec$^{DNN}_{PLDA}$** improves upon **Ivec$^{GMM}_{PLDA}$** by 16% relative EER (from 1.9% to 1.6% absolute) in Cond-all.

5.2. DTW-based SV systems

Table 3 shows the performance of the DTW SV systems (**DTW-MFCC**, **DTW-post^{GMM}**, **DTW-post^{DNN}**, **DTW-onIvec^{GMM}** and **DTW-onIvec^{DNN}**) using different features. All DTW systems outperform the model-based SV approaches in Cond1, with the best performance being achieved by **DTW-post^{DNN}**. However, the DTW systems using MFCCs, DNN-posteriors and GMM-posteriors perform worse than the baseline system in Cond2. As described in Section 3, we address this problem by extracting online i-vectors to be used as feature set for the DTW algorithm.

Experimental results with online i-vectors shown also in Table 3 indicate that **DTW-onIvec^{GMM}** outperforms the baseline **RMAP^{GMM}** by about 50% relative EER (from 5.2% to 2.6% absolute) and 7% relative EER (from 4.1% to 3.8% absolute) for Cond1 and Cond2 respectively. **DTW-onIvec^{DNN}** further improves upon **DTW-onIvec^{GMM}** across all conditions, with gains in relative EER of about 50% (from 2.6% to 1.3% absolute) and 16% (from 3.8% to 3.2% absolute) for Cond1 and Cond2 respectively.

The DTW algorithm plays an important role in achieving good performance for the **DTW-onIvec** systems. We expect that without the sequence matching capability, the online i-vector system, applying an averaging operation instead of preserving the sequential information, would obtain worse results than **DTW-onIvec**. To test this hypothesis, we conducted another experiment by building a SV system (similar to **Ivec$_{PLDA}$**) as follows: A sequence of online i-vectors is extracted and averaged to obtain a representative i-vector of the whole utterance. The PLDA is trained using these averaged online i-vectors as features assuming speaker-phrase as classes. The distance between the enrollment and test speech signal is computed using the PLDA model with the averaged online i-vectors. We built two systems applying this strategy, one with GMM posteriors and another with DNN posteriors referred to as **onIvec$^{GMM}_{PLDA}$** and **onIvec$^{DNN}_{PLDA}$** respectively. Table 3 clearly shows that **onIvec$^{GMM}_{PLDA}$** and **onIvec$^{DNN}_{PLDA}$** perform significantly worse than the **DTW-onIvec** systems. This result highlights the importance of using the DTW algorithm, in addition to the online i-vectors, in obtaining low error rates.

Table 4: Performance of the best baseline and proposed systems on RedDots dataset in terms of EER/minDCF(%).

Systems/Conditions	Cond-all
RMAP^{GMM} (row 1 of Table 1)	1.8/0.36
DTW-onIvec^{DNN} (row 5 of Table 3)	1.3/0.29

The performance of the best baseline (relevance MAP) and proposed (DTW using online i-vector) systems are presented in Table 4 for Cond-all in terms of minDCF. It can be observed that **DTW-onIvec^{DNN}** improves upon **RMAP^{GMM}** by 19% relative minDCF (from 0.36% to 0.29% absolute).

6. CONCLUSIONS

In this paper, we presented relevance MAP, i-vector, JFA and DTW approaches for text-dependent SV task with fixed phrases. The results indicate that the relevance MAP approach performs the best among the baseline systems (which includes the i-vector and JFA) using GMM posteriors. The i-vector and JFA systems largely benefit from using DNNs instead of GMMs for posterior estimation. However the relevance MAP, i-vector and JFA approaches ignore the sequence information of the phonetic units of the phrase. We addressed this problem by using the DTW algorithm exploiting online i-vectors as input features. The proposed DTW approach outperforms the relevance MAP technique by more than 22% relative EER in speaker-mismatch condition. As expected, the proposed approach improves by more than 75% relative EER in content-mismatch condition.

7. ACKNOWLEDGEMENTS

This work was primarily supported by the EU FP7 project "Speaker Identification Integrated Project (SIIP)". This work was also partially supported by the EC H2020 funding, under "Machine Learning of Speech Recognition Models for Controller Assistance (MALORCA)" project.

8. REFERENCES

[1] Daniel Garcia Romero and Carol Y. Espy 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] Petr 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] Anthony Larcher, Kong-Aik Lee, Bin Ma, and Haizhou Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.

[5] Kong Aik Lee, Anthony Larcher, Guangsen Wang, Patrick Kenny, Niko Brümmer, David van Leeuwen, Hagai Aronowitz, Marcel Kockmann, Carlos Vaquero, Bin Ma, et al., "The reddots data collection for speaker recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[6] P. Kenny, T. Stafylakis, P. Ouellet, and M.J. Alam, "Jfa-based front ends for speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 1705–1709.

[7] 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.

[8] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted gaussian mixture models," *Diigital Signal Processing*, vol. 10, pp. 19–41, 2000.

[9] Yun 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.

[10] S. Jelil, R. K. Das, R. Sinha, and S.R. M. Prasanna, "Speaker verification using gaussian posteriorgrams on fixed phrase short utterances," in *Interspeech 2015*, September 2015, pp. 1042–1046.

[11] Vijayaditya Peddinti, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Reverberation robust acoustic modeling using i-vectors with time delay neural networks," *Proceedings of INTERSPEECH. ISCA*, 2015.

[12] Srikanth Madikeri, Ivan Himawan, Petr Motlicek, and Marc Ferras, "Integrating online i-vector extractor with information bottleneck based speaker diarization system," Tech. Rep., Idiap, 2015.

[13] Patrick Kenny, Themos Stafylakis, Pierre Ouellet, and Mohammad Jahangir Alam, "Jfa-based front ends for speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1705–1709.

[14] Patrick Kenny, Themos Stafylakis, J Alam, Pierre Ouellet, and Marcel Kockmann, "Joint factor analysis for text-dependent speaker verification," *Odyssey*, 2014.

[15] Eamonn Keogh and Chotirat Ann Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and information systems*, vol. 7, no. 3, pp. 358–386, 2005.

[16] Michael Brown and L Rabiner, "An adaptive, ordered, graph search technique for dynamic time warping for isolated word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 535–544, 1982.

[17] v. Ramasubramanian, A. Das, and V. P. Kumar, "Text-dependent speaker-recognition using one-pass dynamic programming algorithm," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, May 2006, vol. 1, pp. I–I.

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

[19] Niko Brummer, Lukas Burget, P Kenny, P Matejka, E de Villiers, M Karafiat, M Kockmann, O Glembek, O Plchot, D Baum, et al., "Abc system description for nist sre 2010," *Proc. NIST 2010 Speaker Recognition Evaluation*, pp. 1–20, 2010.

[20] Hossein Zeinali et al., "i-vector/hmm based text-dependent speaker verification system for reddots challenge," in *To appear in InterSpeech 2016*. ISCA, 2016.