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

i-vector/HMM Based Text-dependent Speaker Verification System for RedDots Challenge

Hossein Zeinali [1,2], *Hossein Sameti* [1], *Lukáš Burget* [2]
Jan "Honza" Černocký [2], *Nooshin Maghsoodi* [1] *and Pavel Matějka* [2]

[1] Sharif University of Technology, Tehran, Iran
[2] Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Czech Republic
zeinali@ce.sharif.edu, sameti@sharif.edu, nmaghsoodi@ce.sharif.edu
{burget, cernocky, matejkap}@fit.vutbr.cz

Abstract

Recently, a new data collection was initiated within the RedDots project in order to evaluate text-dependent and text-prompted speaker recognition technology on data from a wider speaker population and with more realistic noise, channel and phonetic variability. This paper analyses our systems built for RedDots challenge – the effort to collect and compare the initial results on this new evaluation data set obtained at different sites. We use our recently introduced HMM based i-vector approach, where, instead of the traditional GMM, a set of phone specific HMMs is used to collect the sufficient statistics for i-vector extraction. Our systems are trained in a completely phrase-independent way on the data from RSR2015 and Libri speech databases. We compare systems making use of standard cepstral features and their combination with neural network based bottle-neck features. The best results are obtained with a score-level fusion of such systems.

Index Terms: text-dependent speaker verification, i-vector, HMM, RedDots challenge

1. Introduction

Speaker verification (SV) is one of challenging areas of speech processing. Historically, text-independent SV has received more attention due to the regular organization of NIST challenges and available of standard data sets. Recently, the R&D in text-dependent SV has gained momentum, and led also to introduction of several standard datasets. RedDots project is one these efforts, where a new data was collected in order to evaluate text-dependent and text-prompted speaker recognition technology on utterances from a wider speaker population and with more realistic noise, channel and phonetic variability. Although not yet completed, the available amount of data already collected was enough to organize a challenge [1] setting the RedDots data as a new text-dependent SV benchmark.

Several methods have been proposed for text-dependent SV, that can be grouped into two categories. The first one includes traditional modeling-scoring methods such as Gaussian Mixture Model–Universal Background Model (GMM-UBM) and Hidden Markov Model–UBM (HMM-UBM) directly producing likelihoods, with new variants mainly elaborating on likelihood ratio scoring [2, 3, 4, 5]. In the second category, new trends in the text-independent SV are ported to the text-dependent case: using mean super-vector for speaker representation and SVM for scoring [6, 7, 8, 9], Joint Factor Analysis with various back-ends [10, 11, 12, 13], i-vector representation with PLDA [14, 15] or Cosine Distance scoring [16, 17], and Gaussian processes [18]. Recently, Deep Neural Network (DNN) have also made it to text-dependent SV [19, 20].

Usually, GMM is used to align frames to Gaussian components when collecting the sufficient statistics for i-vector extraction. The nature of text-dependent speaker verification however encouraged us to propose an *HMM-based approach for performing a better frame alignment*. In [21], we proposed an i-vector/HMM approach, where the Viterbi algorithm is used for the frame alignment. In the first step of this method, a phoneme recognizer is trained and then individual phoneme models are concatenated to create a phrase-specific HMM model for each phrase. These models are then used for extracting sufficient statistics from utterances. It is worth mentioning that although phrase-specific HMM models are used, finally a *single, phrase-independent i-vector extractor* is trained for all phrases exactly the same as the text-independent case. Experimental results on RSR2015 data set [5] have shown that the performance of the proposed HMM-based method is better than other methods especially for rejecting *wrong phrase trials* [21], where the phrase in the test utterance does not match with the enrollment one.

It was proved that the Deep Neural Network (DNN) approaches have a better performance in several speech data mining tasks. Matejka et al. [22] have shown that using bottle-neck DNN features (BN) concatenated to other acoustic features outperformed the DNN method for text-independent SV. In [23], we succeeded in reproducing these results for *text-dependent* task. BN features helped the i-vector/HMM based method, again especially in rejecting the *wrong phrase trials*.

This paper analyses our systems built for RedDots challenge – the effort to collect and compare the initial results on this new evaluation data set obtained at different sites. We describe our experiments on RedDots data with the i-vector/HMM approach. Our systems are trained in a completely phrase-independent way on the data from RSR2015 and Libri speech databases. We compare systems making use of standard cepstral features and their combination with BN features. The best results are obtained with a score-level fusion of such systems.

Our previous work on the i-vector/HMM based method [23] focused on the case of Imposter-Correct non-target trials, where an imposter speaker pronounces the correct phrase. In this paper, we further evaluate the technique on different types of non-

The work was supported by Czech Ministry of Interior project No. VI20152020025 "DRAPAK", and Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science - LQ1602".

http://dx.doi.org/10.21437/Interspeech.2016-1174

target trials. We show that the technique is especially effective in rejecting the target speakers pronouncing a wrong phrase (Target-Wrong trial).

The rest of this paper is organized as follows: in Section 2, the main parts of our system are described. Section 3 presents the experimental setups and Section 4 the results. We conclude in Section 5.

2. i-vector/HMM based approach

To describe our i-vector/HMM based approach, we start with Baum-Welch formula [24]. Let $\mathcal{X} = \{\mathbf{x}_t | t = 1, \dots, T\}$ be the feature vectors from a variable-length input utterance of a specific phrase. Then, the *zero* and *first order statistics* for $\mathcal{X}$; $\mathbf{n}_\mathcal{X} = [N_\mathcal{X}^{(1)}, \dots, N_\mathcal{X}^{(C)}]'$ and $\mathbf{f}_\mathcal{X} = [\mathbf{f}_\mathcal{X}^{(1)'}, \dots, \mathbf{f}_\mathcal{X}^{(C)'}]'$ are given as:

$$N_\mathcal{X}^{(c)} = \sum_t \gamma_t^{(c)} \quad (1)$$

$$\mathbf{f}_\mathcal{X}^{(c)} = \sum_t \gamma_t^{(c)} \mathbf{x}_t \, , \quad (2)$$

where C is the total number of mixture components, and $\gamma_t^{(c)}$ is the posterior (or occupation) probability of frame $\mathbf{x}_t$ being generated by mixture component c. The tuple $\gamma_t = (\gamma_t^{(1)}, \dots, \gamma_t^{(C)})$ is usually referred to as *frame alignment*.

In our HMM-based method, a phoneme recognizer is first trained with 3-state, GMM-based, mono-phone HMMs. This recognizer is the same as in speech recognition. Let F be the total number of mono-phones (i.e. 39), $S = 3F$ be the number of all states, G the number of Gaussian components per state, and $C = SG$ the number of all individual Gaussians, and let (s, g) denote Gaussian component g in state s. Then, for each phrase (based on the transcribed sequence of phonemes in that phrase), a new phrase-specific HMM is constructed by concatenating the corresponding mono-phone HMMs. The Viterbi algorithm is then used to obtain the alignment of the frames to the HMM states, and within each state s, the GMM alignment $\gamma_t^{(s,g)}$ is computed for each frame t. We can now re-interpret the pair (s, g) as one out of C Gaussians and we can substitute $\gamma_t^{(c)}$ in Eqs. (1) and (2) by $\gamma_t^{(s,g)}$, so that the zero and first order statistics can be written as:

$$\mathbf{n}_\mathcal{X} = [N_\mathcal{X}^{(1,1)}, \dots, N_\mathcal{X}^{(s,g)}, \dots, N_\mathcal{X}^{(S,G)}]'$$
$$\mathbf{f}_\mathcal{X} = [\mathbf{f}_\mathcal{X}^{(1,1)'}, \dots, \mathbf{f}_\mathcal{X}^{(s,g)'}, \dots, \mathbf{f}_\mathcal{X}^{(S,G)'}]' \, ,$$

where:

$$N_\mathcal{X}^{(s,g)} = \sum_t \gamma_t^{(s,g)} \quad (3)$$

$$\mathbf{f}_\mathcal{X}^{(s,g)} = \sum_t \gamma_t^{(s,g)} \mathbf{x}_t \, , \quad (4)$$

Note that in Eqs. (3) and (4), due to the typically short duration of phrases, not all phonemes are used in the phrase-specific HMM. Therefore the alignment of frames to the Gaussian components is often sparse and most of the $\gamma_t^{(s,g)}$ values are zero.

When comparing this HMM structure with GMM, we should note that it has two main advantages. The first one is better localized frame alignment to Gaussian components — posteriors of each frame are non-zero in exactly G components. The second advantage is high chance of rejecting *wrong phrase trials*; for these, the phoneme sequence of the input utterance does not match with the phonemes of the pass-phrase. The frame posteriors and consequently zero and first-order statistics are wrong. In this case, the extracted i-vector is very different from enrollment i-vectors and can be easily rejected.

3. Experimental setups

3.1. Data set

The current snapshot of RedDots used for this challenge contains 62 speakers including 49 males and 13 females. 41 speakers are used as target speakers (35 males and 6 females) and the other ones are considered as unseen imposters. RedDots dataset consists of four subsets: In Part-01, each speaker uttered 10 common phrases, so for this part, three types of non-target trials can be considered: *Imposter-Correct*, *Imposter-Wrong* (i.e. trials corresponding to an imposter speaker in the test utterance pronouncing correct or wrong phase, respectively) and *Target-Wrong* trials. We report results separately for tree conditions corresponding to the three non-target trial types. All three conditions share the same *Target-Correct* trials. The results are reported in terms of Equal Error Rate (EER) and Normalized Detection Cost Function as defined for NIST SRE08 ($\mathrm{NDCF}_{\mathrm{old}}^{\mathrm{min}}$) and NIST SRE10 ($\mathrm{NDCF}_{\mathrm{new}}^{\mathrm{min}}$).

In Part-02, each speaker pronounced 10 unique phrases and in Part-03, two free-choice phrases were used. In Part-02 and -03, there are no Imposter-Correct trials. The last part of this data set (i.e. Part-04) is the combination of previous three parts. For Part-04, there are two different tasks (i.e. text-dependent and text-prompted). We dealt with the text-dependent task only.

For UBM and i-vector extractor training, the combination of the RSR2015 dataset Part-1 [5] and 100 hours of Libri speech [25] was used. The RSR2015 dataset comprises recordings from 157 male and 143 female speakers each pronouncing 30 different phrases from TIMIT in 9 distinct sessions. The *Train-Clean-100* part of Libri speech includes 251 speakers (126 males and 125 females). Switchboard data was used to train of DNN for extracting BN features.

3.2. Features

We have experimented with several different configurations for extracting the standard cepstral features. For the experiments reported in this paper, we have selected 60-dimensional PLP and MFCC features, both extracted from 16 kHz signal using HTK [26] with a similar configuration: 25 ms Hamming-windowed frames with 15 ms overlap. For each utterance, the features are normalized using cepstral-mean and -variance normalization after dropping the initial and final silence frames.

In addition to the cepstral features, 80-dimensional DNN based stacked bottle-neck features are used [27, 22, 23]. Note that these features are extracted from data down-sampled to 8 kHz as, at the time of running these experiments, the only BN DNN available was trained on 8 kHz conversational telephone Switchboard data. In the experiments presented in this paper, we never use the BN alone. Instead, we concatenate BN features and MFCC features to form 140-dimensional MFCC+BN features. Such features proved to provide superior performance in our former experiments [22, 23].

3.3. System configuration

HMM with 3 states and 8 Gaussian components for each of 39 mono-phones were used for the alignment (resulting in total number of 936 Gaussian components). All reported results are obtained with the i-vector/HMM based systems. The 400- and 600-dimensional i-vectors are extracted using gender-dependent HMM-UBMs and i-vector extractors. Cosine distance is used to obtain speaker verification scores.

Due to the very short utterances with variable content, we were not able to successfully apply any channel compensation.

Table 1: *Results on Part-01 males with MFCC features comparing different approaches and i-vector dimensionalities.*

Method	Trial type	EER [%]	$\text{NDCF}^{\text{min}}_{\text{old}}$	$\text{NDCF}^{\text{min}}_{\text{new}}$
Relevance MAP/GMM	Imp-Corr	1.98	0.0848	0.2879
	Tar-Wrg	4.01	0.1733	0.4960
	Imp-Wrg	**0.34**	0.0135	0.0488
i-vector/GMM (dim: 600)	Imp-Corr	2.07	0.0899	0.3105
	Tar-Wrg	3.76	0.1762	0.4275
	Imp-Wrg	0.43	0.0153	0.0435
i-vector/HMM (dim: 400)	Imp-Corr	2.31	0.0893	**0.2250**
	Tar-Wrg	1.30	0.0580	0.1423
	Imp-Wrg	0.56	0.0121	0.0404
i-vector/HMM (dim: 600)	Imp-Corr	**1.88**	**0.0809**	0.2271
	Tar-Wrg	**1.11**	**0.0338**	**0.0509**
	Imp-Wrg	0.46	**0.0106**	**0.0228**

We also tried different score normalization methods, but most of them failed. Therefore, all reported results are without any normalization. Note that in [21, 23], we used phrase-dependent regularized WCCN and phrase-dependent S-Norm. We have also shown there that if channel compensation and normalization were trained on an independent dataset of different phrases (i.e. similarly as in text-independent case), they failed. It seems that in text-dependent SV, and especially with our HMM based method, we cannot do any phrase independent normalization. For score fusion, we used simple score averaging with equal weights for each sub-system.

4. Results

4.1. Comparison to the traditional approaches

To justify our choice of the i-vector/HMM approach for the RedDots challenge, we compare it to two simpler and more conventional baseline methods. In Table 1, *Relevance MAP/GMM* corresponds to the traditional approach based on Relevance MAP adaptation of GMM-UBM (1024 components) and log-likelihood ratio scoring [2], which is still the standard approach to text-dependent speaker recognition. The second baseline system *i-vector/GMM* is based on the standard i-vector approach and the only difference from our proposed *i-vector/HMM* system is the use of GMM-UBM (1024 components) rather than HMM-UBM. As can be seen, the i-vector/HMM system outperforms (or at least performs comparably to) the baselines for all conditions. The improvement is especially significant for the Target-Wrong condition, which shows that the approach is very effective in verifying the correctness of the phrase.

In almost all cases (and especially for the wrong phrase conditions), the performance of 600-dimensional i-vectors is better than 400-dimensional ones. Therefore, we continue reporting only the results with 600-dimensional i-vectors.

4.2. Feature comparison

In Table 2, we compare the performance of the systems making use of the different features. We observe that, compared to the cepstral features, the MFCC+BN features are much better in rejecting *wrong phrase trials*. However, the price paid for it is considerably worse performance on *Imposter-Correct* trials. In order to show the cause of this behavior, the score distributions of different trial types for MFCC and MFCC+BN are plotted in Figure 1.

Table 2: *Results on Part-01 males comparing different features for different trial types.*

Features	Trial type	EER [%]	$\text{NDCF}^{\text{min}}_{\text{old}}$	$\text{NDCF}^{\text{min}}_{\text{new}}$
MFCC	Imp-Corr	**1.88**	0.0809	**0.2271**
	Tar-Wrg	1.11	0.0338	0.0509
	Imp-Wrg	0.46	0.0106	0.0228
PLP	Imp-Corr	2.13	**0.0738**	0.2339
	Tar-Wrg	1.20	0.0373	0.0759
	Imp-Wrg	0.49	0.0129	0.0327
MFCC+BN	Imp-Corr	2.92	0.1295	0.4246
	Tar-Wrg	**0.37**	**0.0081**	**0.0154**
	Imp-Wrg	**0.22**	**0.0036**	**0.0059**

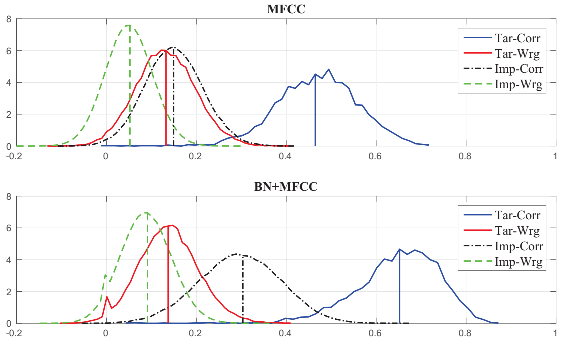

Figure 1: *Score distribution of different trial types of male trials for MFCC and MFCC concatenated to Bottle-Neck (i.e. MFCC+BN). The vertical lines show the mean of normal distribution fitted to the scores.*

We can see that the distributions of correct phrase trials (for both target and imposter speakers) move considerably to the right (i.e. scores higher) in the case of MFCC+BN features as compared to the cepstral features. This indicates that i-vectors extracted from MFCC+BN features contain also significant amount of information about phonetic content of the phrases, which makes i-vectors extracted from the same phrase more similar. This is understandable as the BN features were trained capture information important for phone recognition. As the distribution of correct phrase trials moves away from the wrong phrase trials in the case of the MFCC+BN features, it makes it easy to reject the wrong phrase trials. On the other hand, the overlap between *Target-Correct* and *Imposter-Correct* scores distributions slightly increases for the MFCC+BN features, which results in a poorer performance for *Imposter-Correct* condition. In order to take advantage of the excellent performance of MFCC+BN features on wrong phrase trials and better performance of cepstral features on *Imposter-Correct* trials, we fuse scores of systems based on these different features.

Note that performance of the MFCC+BN features on RedDots is contradicting our findings on the RSR2015 dataset [23]. On RSR2015, the performance of MFCC+BN was better than MFCC also for *Imposter-Correct* condition. In our RSR2015 experiments, the data for training HMM-UBM and i-vector extractor contained the same phrase as the enrollment and test utterances. However, this is not the case for most of the phrased in RedDots. Therefore, we suspected that the BN features might be sensitive to such mismatch between the training and test data. To prove hypothesis, we concentrate on three out of

Table 3: *The final results for all 4 parts of the RedDots challenge (fusion of three systems with different features). The gray cells correspond to the "unreliable" conditions, for which we make less than 30 errors of type I or II (See Doddington's rule of 30 [28]).*

		Male			**Female**		
Part	Non-target trial type	EER [%]	$\mathrm{NDCF}_{\mathrm{old}}^{\mathrm{min}}$	$\mathrm{NDCF}_{\mathrm{new}}^{\mathrm{min}}$	EER [%]	$\mathrm{NDCF}_{\mathrm{old}}^{\mathrm{min}}$	$\mathrm{NDCF}_{\mathrm{new}}^{\mathrm{min}}$
Part-01	Imposter-Correct	1.60	0.0615	0.2132	2.75	0.1156	0.1877
	Target-Wrong	0.49	0.0105	0.0120	0.63	0.0174	0.0174
	Imposter-Wrong	0.25	0.0039	0.0056	0.32	0.0111	0.0221
Part-02	Target-Wrong	0.34	0.0055	0.0219	0.17	0.0055	0.0069
	Imposter-Wrong	0.28	0.0035	0.0038	0.17	0.0017	0.0017
Part-03	Target-Wrong	0.16	0.0016	0.0016	1.52	0.0152	0.0152
	Imposter-Wrong	0.16	0.0016	0.0016	0.76	0.0076	0.0076
Part-04	Imposter-Correct	1.35	0.0542	0.1910	2.23	0.0899	0.1248
	Target-Wrong	0.19	0.0033	0.0232	0.36	0.0085	0.0152
	Imposter-Wrong	0.19	0.0022	0.0034	0.27	0.0073	0.0152

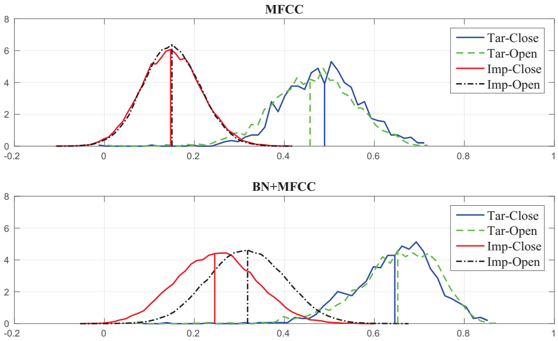

Figure 2: *Score distributions of two different phrase sets for correct male trials for MFCC and MFCC+BN. The vertical lines show the means of normal distributions fitted to scores.*

the ten phrases from Part-01 of RedDots that are common with RSR2015 (as mentioned earlier, RSR2015 is used for training in our RedDots experiments). In Figure 2, we plot correct trials score distributions for two separate phrase sets: *"Close"* with just the three common phrases, and *"Open"* with all other phrases. It is clear that for MFCC+BN features, the target and non-target scores distributions of the Close set are farther from each other than for the Open set. This shows that we can expect much better performance of the BN features when test phrases are included in training data.

Another reason for the worse performance of MFCC+BN features could be their higher dimensionality compared to MFCCs (140 vs 60). However, any our attempt to reduce the dimensionality of the MFCC+BN features only resulted if a further performance degradation.

4.3. Final fusions

The final results were obtained from the fusion of three i-vector/HMM based system, each making use one of the three different features: MFCC, PLP and MFCC+BN. For score fusion, we simply averaged the scores with the same weights. We experimented with a trained logistic regression fusion, but its performance was as good as the simple averaging. For four parts of the RedDots challenge, the results are summarized in Table 3. In this table, the gray cells show results that should be taken with care — operating points are placed on very steep regions of DET curves and a little change in the threshold change them considerably.

Comparing Part-01 results in Table 3 and Table 2 shows that the fusion often performs significantly better than any of the individual systems while it never performs significantly worse. According to in Table 3, most of the wrong trials received low scores, which, we believe, is due to using our HMM method and BN features: when Viterbi force alignment is used for frame alignment with different-phrase HMM, most of the frames are assigned to incorrect Gaussian components, the calculated posteriors are wrong and consequently this type of trials can be easily rejected.

The results on Part-03 deserve a comment: In this part, each speaker was free to select two pass-phrases. Unfortunately some of the speakers used non-English words in the pass-phrase and so we had to map non-English phonemes to the nearest English ones by hand.

5. Conclusions

This paper describes our system for the RedDots challenge data. Our system builds on our i-vector/HMM based method for text-dependent SV where a new HMM structure is used for UBM modeling instead of traditional GMM. Generally, we have seen that in text-dependent speaker verification, we cannot do any channel compensation and score normalization as it is usual in text-independent systems. This happens due to very short utterances, where phonetic variation is dominant compared to the speaker and channel variation and so the i-vectors of different phrases have large distances from each other.

The investigation into features has shown that the performance of the cepstral features is better than MFCC+BN ones for imposter-correct trials, but in wrong trials, the performance of BNs is much better. We have also confirmed that in the close phrase-set task (i.e. when test phrases are in present the training data) BN features work better compared to the open phrase-set task. Our best result for the RedDots challenge was achieved by fusing three systems based on different features in score domain.

Comparing 400- and 600-dimensional i-vectors showed that larger i-vector is better for wrong trials and that the performance does not change much for correct trials. This suggests that larger i-vectors can better represent both the speaker and the phrase.

6. References

[1] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, D. van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma *et al.*, "The RedDots data collection for speaker recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

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

[3] D. T. Toledano, C. Esteve Elizalde, and J. Gonzalez-Rodriguez, "Phoneme and sub-phoneme t-normalization for text-dependent speaker recognition," in *Odyssey-The Speaker and Language Recognition Workshop*. International Speech Communication Association, 2008.

[4] A. Larcher, J.-F. Bonastre, and J. S. Mason, "Constrained temporal structure for text-dependent speaker verification," *Digital Signal Processing*, vol. 23, no. 6, pp. 1910–1917, 2013.

[5] 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.

[6] C. Dong, Y. Dong, J. Li, and H. Wang, "Support vector machines based text dependent speaker verification using HMM supervectors," in *Odyssey-The Speaker and Language Recognition Workshop*, 2008, p. 31.

[7] S. Novoselov, T. Pekhovsky, A. Shulipa, and A. Sholokhov, "Text-dependent GMM-JFA system for password based speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 729–737.

[8] H. Aronowitz, "Text dependent speaker verification using a small development set," in *Odyssey-The Speaker and Language Recognition Workshop*, 2012.

[9] H. Sun, K. A. Lee, and B. Ma, "A new study of GMM-SVM system for text-dependent speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4195–4199.

[10] P. Kenny, T. Stafylakis, J. Alam, P. Ouellet, and M. Kockmann, "Joint factor analysis for text-dependent speaker verification," *Odyssey-The Speaker and Language Recognition Workshop*, 2014.

[11] 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*. IEEE, 2014, pp. 1705–1709.

[12] T. Stafylakis, P. Kenny, J. Alam, and M. Kockmann, "JFA for speaker recognition with random digit strings," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[13] T. Stafylakis, P. Kenny, M. J. Alam, and M. Kockmann, "Speaker and channel factors in text-dependent speaker recognition," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 24, no. 1, pp. 65–78, 2016.

[14] T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann, and P. Dumouchel, "Text-dependent speaker recognition using PLDA with uncertainty propagation," in *Interspeech*, 2013, pp. 3684–3688.

[15] A. Larcher, K. A. Lee, B. Ma, and et al., "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.

[16] H. Zeinali, E. Kalantari, H. Sameti, and H. Hadian, "Telephony text-prompted speaker verification using i-vector representation," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4839–4843.

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

[18] N. Maghsoodi, H. Sameti, and H. Zeinali, "Localized discriminative Gaussian process latent variable model for text-dependent speaker verification," in *ESANN*. IEEE, 2016.

[19] E. Variani, X. Lei, E. McDermott, I. Lopez Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4052–4056.

[20] 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.

[21] H. Zeinali, H. Sameti, and L. Burget, "HMM-based phrase-independent i-vector extractor for text-dependent speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, in.press.

[22] P. Matejka, O. Glembek, O. Novotny, O. Plchot, F. Grezl, L. Burget, and J. Cernocky, "Analysis of DNN approaches to speaker identification," in *ICASSP*, 2016.

[23] H. Zeinali, L. Burget, H. Sameti, O. Glembek, and O. Plchot, "Deep neural networks and hidden Markov models in i-vector-based text-dependent speaker verification," in *Odyssey-The Speaker and Language Recognition Workshop*, 2016.

[24] 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, 2011.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.

[26] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, *The HTK book*. Entropic Cambridge Research Laboratory Cambridge, 1997, vol. 2.

[27] P. Matejka, L. Zhang, T. Ng, H. Mallidi, O. Glembek, J. Ma, and B. Zhang, "Neural network bottleneck features for language identification," *Odyssey-The Speaker and Language Recognition Workshop*, pp. 299–304, 2014.

[28] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The {NIST} speaker recognition evaluation overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, no. 23, pp. 225 – 254, 2000.