

Phone-Centric Local Variability Vector for Text-Constrained Speaker Verification

Liping Chen¹, Kong Aik Lee², Bin Ma², Wu Guo¹, Haizhou Li², and Li Rong Dai¹

¹National Engineering Laboratory for Speech and Language Information Processing, USTC, China

²Institute for Infocomm Research, A*STAR, Singapore

clp2011@mail.ustc.edu.cn, kalee@i2r.a-star.edu.sg

Abstract

This paper investigates the use of frame alignment given by a deep neural network (DNN) for text-constrained speaker verification task, where the lexical contents of the test utterances are limited to a finite set of vocabulary. The DNN makes use of information carried by the target and its contextual frames to assign it probabilistically to one of the phonetic states. The frame alignment is therefore more precise and less ambiguous than that generated by a Gaussian mixture model (GMM). Using the DNN alignment, we show that an i-vector can be decomposed into segments of local variability vectors, each corresponding to a monophone, where each local vector models session variability given the phonetic context. Based on the local vectors, the content matching between the utterances for comparison can be accomplished in the PLDA scoring. Experiments conducted on the RSR2015 database shows that the proposed phone-centric local variability vector achieves a better performance compared to the i-vector.

Index Terms— text-constrained speaker verification, deep neural network, phone-centric local variability

1. Introduction

Over recent years, many approaches based on *Gaussian mixture model-universal background model* (GMM-UBM) [1] have been proposed for speaker verification [2]. Based upon similar framework as the eigenvoice model [3], the i-vector was proposed in [4] and soon became the mainstream method in speaker verification. Similar to a supervector [5], an i-vector is a fixed-length representation of a speech utterance, which typically consists of varying number of frames. Besides, an i-vector offers a low-dimensional representation of the total variability, containing both speaker and channel information, rendered in an utterance. Channel compensation could then be performed using the *probabilistic linear discriminant analysis* (PLDA) [6], which also serves as the backend classifier.

I-vectors have proven to be very effective for text-independent verification when both the training and test utterances are of long duration. In text-constrained speaker verification task, the lexical content of the spoken utterances are confined to a finite known set and always of short duration [7]. For short utterances, the content mismatch between the enrollment and test utterances makes it difficult for i-vectors to be applied directly [8, 9]. In [11], we proposed the *local variability model* (LVM) with the aim to capture the local variability associated with individual Gaussians of the UBM. It unifies a shared i-vector among all the Gaussians by assigning a local variability

vector to each Gaussian component. LVM offers a flexible way to compare two utterances in a content-wise manner which may provide a viable solution to the content mismatch problem between utterances of short duration.

Individual Gaussian components of the UBM can be thought of the representation of some low-level acoustic events [10]. However, it is generally difficult to associate such events to any well-defined phonetic classes since the UBM is trained in an unsupervised manner. The lexical content modeled by individual Gaussians is therefore ambiguous without clear phonetic definition. This problem was addressed in [12] and [13], using *deep neural network* (DNN) for i-vector extraction. With the DNN for triphone models trained for *automatic speech recognition* (ASR), speech frames are aligned to a predefined set of tied-states, which are treated as Gaussian-like units in the i-vector framework. Thus, every unit is endowed with clear phonetic meaning.

In [14], a method to perform the content-matching at the statistics level was proposed, which was shown to be effective for the tasks where the lexical contents of the test utterance is confined to those for enrolment. However, in such a method, the i-vectors need to be estimated for the enrolment and test utterances in every trial, resulting in complex computation in the testing phase. In this paper, treating the output nodes of a DNN trained as a monophonic acoustic model for ASR as Gaussian-like units, we propose to estimate the phone-centric local variability vectors. The benefit of such phone-centric local vectors is that, they are estimated independently before speaker comparison, and can be selected to be compared according to their corresponding phonetic contents during the subsequent PLDA scoring, performing content-matching. We show that such local vectors outperform the (DNN) i-vector in the content-matching.

The rest of this paper is organized as follows. Section 2 reviews the i-vector extraction briefly. Section 3 describes the local variability modeling. Section 4 proposes the phone-centric local variability vectors. Section 5 shows the results of our experiments. Section 6 is the conclusion.

2. I-vector extraction with GMM vs. DNN

2.1 GMM i-vector

The purpose of i-vector extraction is to find a compressed representation of the speaker and channel information rendered in a variable-length utterance [4]. The fundamental assumption is that the feature vector sequence of the utterance is generated from a session-specific GMM. Furthermore, the mean supervector of the GMM \mathbf{m}_r is confined to a low-dimensional subspace \mathbf{T} with origin $\boldsymbol{\mu}$, as follows

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

10.21437/Interspeech.2015-90

where $r = 1, 2, \dots, R$ denotes the session index. The matrix \mathbf{T} is referred to as the total variability matrix, and the model is referred to as the total variability model. The latent variable \mathbf{w}_r is session-specific whose posterior mean is the so-called i-vector.

Let C be the number of Gaussians in the UBM, (1) can be decomposed as

$$\mathbf{m}_{c,r} = \boldsymbol{\mu}_c + \mathbf{T}_c \mathbf{w}_r, \text{ for } c = 1, 2, \dots, C, \quad (2)$$

where $\boldsymbol{\mu}_c$ and \mathbf{T}_c are the mean vector and loading matrix associated with the c -th Gaussian of the UBM. Let $\mathcal{O} = \{o_1, o_2, \dots, o_N\}$ represents the feature sequence of a given utterance. An i-vector is computed as

$$\boldsymbol{\phi} = \mathbf{L}^{-1} \left[\sum_{c=1}^C \mathbf{T}_c^T \tilde{\mathbf{F}}_c \right], \quad (3)$$

where \mathbf{L} is the posterior precision given by

$$\mathbf{L} = \mathbf{I} + \sum_{c=1}^C N_c \mathbf{T}_c^T \mathbf{T}_c. \quad (4)$$

In the above equations, $\{N_c, \tilde{\mathbf{F}}_c\}$ are the utterance-dependent Baum-Welch statistics computed based on the UBM. In particular, N_c is the zero-order statistics computed for the c -th Gaussian by summing the frame occupancy $\gamma_c(t)$ over the entire sequence, as follows

$$N_c = \sum_t \gamma_c(t), \quad (5)$$

while

$$\tilde{\mathbf{F}}_c = \boldsymbol{\Sigma}_c^{-1/2} \left[\sum_t \gamma_c(t) (o_t - \boldsymbol{\mu}_c) \right] \quad (6)$$

is the first-order statistics centred to the mean $\boldsymbol{\mu}_c$ and whiten with respect to covariance $\boldsymbol{\Sigma}_c$ of the UBM [15]. For brevity, we dropped the session index r in (3) to (6).

2.2 DNN i-vector

In the state-of-the-art ASR systems, context-dependent phones (e.g., triphones) serve as the basic units that are represented by a number of hidden Markov states. Typically, the Markov states are tied across the context-dependent phones using a decision tree. In the conventional GMM-HMM framework, the emission probabilities of the tied-states are modelled with GMMs [16]. In a DNN-HMM hybrid system, the observation probabilities are estimated through a DNN. More precisely, each output neuron of the DNN is trained to estimate the posterior probability of tied-states given the acoustic observations [17].

By treating the set of tied-states $\mathcal{S} = \{s_1, s_2, \dots, s_L\}$ modeled by the DNN as Gaussian-like units in the UBM, it was shown in [12] and [13] that the Baum-Welch statistics, and therefore the i-vector, could be extracted by replacing the frame alignment with the tied-states posterior, as follows

$$\gamma_l(t) \leftarrow p(s_l | \mathbf{x}_t), \quad (7)$$

where $l = 1, 2, \dots, L$ and \mathbf{x}_t is the feature vector used for DNN. The remaining operations follow as if the UBM consists of $C = L$ Gaussian-like units. It is worth mentioning that the feature vectors used for speaker modeling can be different from those used for the DNN. The latter usually occupies a longer window (typically 10 to 15 frames) and is richer in phonetic information.

3. Local variability model

Figure 1 shows the graphical model of an i-vector extractor. Our

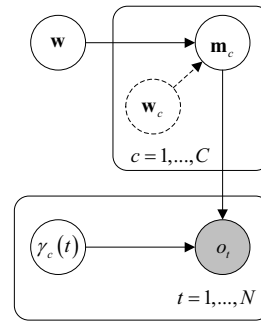


Fig. 1: Graphical model illustrating the difference between the total and local variability models. The latent variable \mathbf{w} is used for total variability model, while \mathbf{w}_c , for $c = 1, 2, \dots, C$, are for local variability model. For brevity, we dropped the reference to session index r .

aim is to model the speaker and channel variability jointly with the latent variable \mathbf{w}_r . This is achieved by tying the latent variable across the observed frames and Gaussians (or Gaussian-like units when DNN is used for frame alignment). Firstly, the variable \mathbf{w}_r is tied across all the observed frames in a given utterance based on the assumption that the speaker and recording channel are kept consistent throughout an utterance. Secondly, the latent variable \mathbf{w}_r is shared among all the Gaussians. The second point is also reflected in (2), where the same variable \mathbf{w}_r is shared among all $c = 1, 2, \dots, C$. I-vector is extremely effective for long utterances. However, it suffers from the problem of content mismatch when the test utterances are of short duration.

Also shown in Fig. 1 is the local variability model (LVM) [11], where the latent variable $\mathbf{w}_{c,r}$ is now moved inside the rectangular box. This relaxes the tying across Gaussians, where one latent variable is assigned to each Gaussian, which can be described in mathematical form as:

$$\mathbf{m}_{c,r} = \boldsymbol{\mu}_c + \mathbf{T}_c \mathbf{w}_{c,r}. \quad (8)$$

The posterior means of the latent variables $\{\mathbf{w}_{c,r}\}_{c=1}^C$ collectively form the local variability vector. An i-vector models the variability across the entire utterance. The local variability vector captures the information dedicated to individual Gaussian. By so doing, two utterances can be compared in a content-wise manner, which provides a solution to the content mismatch problem when the test utterances are of short duration.

Our previous results in [11] show that the LVM suffers from the inaccurate frame alignment based on the UBM. In this paper, we show that this issue could be alleviated by the use of frame alignment from DNN. It is generally conceded that the benefit of using the DNN posterior is largely due to the strong representation abilities of DNN. Furthermore, each phone state has a clear phonetic definition as opposed to the Gaussian components in the UBM which are mostly generic to all phones. In particular, the outputs associated with the phone states could be grouped to form phone or word units, as we shall see in the next section.

4. Phone-centric local variability vector

4.1. Monophonic acoustic model

In this paper, we use the acoustic model of monophones instead of triphones. Fig. 2 shows the structure of the DNN. The phones /f/ and /tcl/ are taken as the examples for illustrative purpose. Each phone is modeled with three emitting states, i.e., s_2, s_3 and s_4 . The state set \mathcal{S} from all the monophones are modeled by

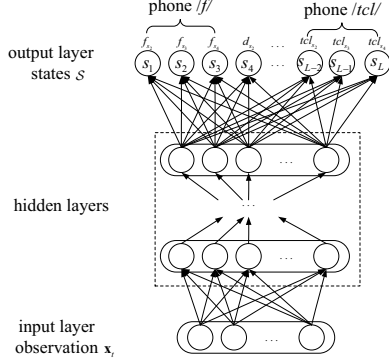


Figure 2: The deep neural network used in a monophonic acoustic model in automatic speech recognition.

the output nodes of the DNN. Given an input observation \mathbf{x} , (usually concatenated with its contextual frames), the values of the output nodes are the state posterior probability $p(s_i|\mathbf{x}_r)$.

In our implementation, the feature vectors for the DNN are 120-dimensional filter bank coefficients with first and second derivatives appended. Every current frame is concatenated with the 5 frames before and after it. We adopted the dictionary from TIMIT dataset, which consists of 61 phones. With each monophone having 3 Markov states, a total of $61 \times 3 = 183$ states need to be modeled. The structure of the network is $1320-(2048 \times 5)-183$.

4.2. Phone-centric local vector estimation

In the estimation of local vectors, instead of assigning a latent variable to each Gaussian-like unit as in [11], a latent variable is shared by the units of the same phone, leading the model to be:

$$\mathbf{m}_{k,r} = \boldsymbol{\mu}_k + \mathbf{V}_k \mathbf{w}_{k,r}, \quad (9)$$

Here, we decompose the mean supervector \mathbf{m}_r for each session r into $K = 61$ groups, where $\mathbf{m} = [\mathbf{m}_{1,r}^T, \dots, \mathbf{m}_{K,r}^T]^T$ and $k = 1, \dots, K$. Each group is made up of three units belonging to the same phone. Following the same grouping, $\boldsymbol{\mu}_k$, \mathbf{V}_k and $\mathbf{w}_{k,r}$ are the global mean vector, loading matrix, and latent variable associated with the k -th phone.

For a given utterance, Baum-Welch statistics are estimated first according to the frame alignment $\gamma_{l,r}(t)$ given by the state posterior $p(s_i|\mathbf{x}_r)$. Let, $N_{l,r}$ be the frame occupancy count to the l -th unit, we form the following matrix containing the zero-order statistics from all units:

$$\boldsymbol{\Gamma}_r = \begin{bmatrix} N_{1,r} \mathbf{I} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & N_{2,r} \mathbf{I} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & N_{L,r} \mathbf{I} \end{bmatrix}, \quad (10)$$

where \mathbf{I} is the identity matrix of the same size as the feature vector o_i for speaker modeling. The first-order statistics $\tilde{\mathbf{F}}_r$ is centered to the global mean vectors and whitened with the covariance matrices in a unit-wise manner. The phone-centric local variability vectors are computed as

$$\phi_{k,r} = \mathbf{L}_{k,r}^{-1} \mathbf{V}_k^T \tilde{\mathbf{F}}_{k,r} \quad (11)$$

$$\mathbf{L}_{k,r}^{-1} = [\mathbf{I} + \boldsymbol{\Gamma}_{k,r} \mathbf{V}_k^T \mathbf{V}_k]^{-1} \quad (12)$$

for $k = 1, \dots, K$. Here, $\tilde{\mathbf{F}}_{k,r}$ and $\boldsymbol{\Gamma}_{k,r}$ are the statistics on the units of the k -th phone. Also, $\boldsymbol{\Gamma}_{k,r}$ is a diagonal matrix containing the zero-order statistics of the 3 units pertaining to the k -th phone. The set of local variability vectors $\{\phi_{k,r}\}_{k=1}^K$ represent the phone-centric speaker and channel variability in an utterance, and their concatenation can be used as an alternative i-vector.

4.3. Model estimation

EM algorithm [18] is applied to train the phone-centric local variability model. The E-step is the posterior inference given by (11) and (12). For model estimation, the auxiliary function is given by:

$$\mathcal{Q} = \sum_{r=1}^R \sum_{k=1}^K \mathbb{E} \left\{ \mathbf{w}_{k,r}^T \mathbf{V}_k^T \tilde{\mathbf{F}}_{k,r} - \frac{1}{2} \boldsymbol{\Gamma}_{k,r} \mathbf{w}_{k,r} \mathbf{w}_{k,r}^T \right\}. \quad (13)$$

Taking the derivative of \mathcal{Q} with respect to \mathbf{V}_k and set resulting expression to zero, we obtain the M-step as follow:

$$\mathbf{V}_k \sum_{r=1}^R \boldsymbol{\Gamma}_{k,r} \mathbb{E} [\mathbf{w}_{k,r} \mathbf{w}_{k,r}^T] = \sum_{r=1}^R \tilde{\mathbf{F}}_{k,r} \mathbb{E} [\mathbf{w}_{k,r}^T]. \quad (14)$$

4.4. PLDA scoring for content-matching

Via concatenation, a variability vector is formed for a given utterance as $\phi_r = [\phi_{1,r}^T, \dots, \phi_{K,r}^T]^T$, on which a PLDA model can be trained to perform channel compensation. With the PLDA model, the marginal probability of the vector ϕ_r will be:

$$p(\phi_r) = \mathcal{N}(\boldsymbol{\omega}, \Psi \Psi^T + \Phi \Phi^T + \Lambda) \quad (15)$$

where $\boldsymbol{\omega}$ is the global mean vector; Ψ and Φ are the speaker and channel subspaces; Λ is the diagonal residual covariance. Due to the possible content mismatch between the compared utterances, during the PLDA scoring, the independent estimation of the local vectors offers a chance for them to be selected for comparison, leading the common phones in both the enrolment and test utterances to be scored with the PLDA model.

Mathematically, Ψ can be decomposed with respect to individual local vectors as $\Psi = [\Psi_1^T, \dots, \Psi_K^T]^T$, so do $\boldsymbol{\omega}$, Φ and Λ . For any vector concatenated with M vectors chosen from the K local vectors ($M \leq K$) in ϕ_r , denoted as $\phi_r = [\phi_{i_1,r}^T, \dots, \phi_{i_M,r}^T]^T$ with i_m to be the indices of the selected local vectors in ϕ_r for $m = 1, \dots, M$, its marginal probability is:

$$p(\phi_r) = \mathcal{N}(\tilde{\boldsymbol{\omega}}, \mathbf{Z}), \quad (16)$$

where the mean vector is $\tilde{\boldsymbol{\omega}} = [\boldsymbol{\omega}_{i_1}^T, \dots, \boldsymbol{\omega}_{i_M}^T]^T$ and the covariance matrix \mathbf{Z} composed as:

$$\mathbf{Z} = \begin{bmatrix} \Psi_{i_1} \Psi_{i_1}^T + \Phi_{i_1} \Phi_{i_1}^T + \Lambda_{i_1} & \dots & \Psi_{i_1} \Psi_{i_M}^T + \Phi_{i_1} \Phi_{i_M}^T \\ \vdots & \ddots & \vdots \\ \Psi_{i_M} \Psi_{i_1}^T + \Phi_{i_M} \Phi_{i_1}^T & \dots & \Psi_{i_M} \Psi_{i_M}^T + \Phi_{i_M} \Phi_{i_M}^T + \Lambda_{i_M} \end{bmatrix} \quad (17)$$

In (17), the subscript i_m indicates the subcomponents in the model parameters corresponding to the i_m -th local vectors for $m = 1, \dots, M$. Specifically, Λ_{i_m} is the i_m -th block of Λ on the diagonal. For brevity, the model parameters in (17) can be integrated as $\tilde{\Psi} = [\Psi_{i_1}^T, \dots, \Psi_{i_M}^T]^T$, $\tilde{\Lambda} = \text{diag}(\text{diag}(\Lambda_{i_1}); \dots; \text{diag}(\Lambda_{i_M}))$ and $\tilde{\Phi} = [\Phi_{i_1}^T, \dots, \Phi_{i_M}^T]^T$ [19], leading the covariance matrix \mathbf{Z} in (16) to be $\mathbf{Z} = \tilde{\Psi} \tilde{\Psi}^T + \tilde{\Phi} \tilde{\Phi}^T + \tilde{\Lambda}$.

In the scoring phase, given the variability vectors estimated from the enrolment and test utterances ϕ_e , ϕ_t and the set of indices of the phones to be compared $\{i_m | m=1, \dots, M\}$, the con-

Table I: Performance in terms of EER(%) and minDCF08($\times 100$) comparing three methods: GMM i-vector, DNN i-vector, phone-centric variability vector and selected phone-centric local vectors, denoted as 'i-vec', 'DNN i-vec', 'ph-vec' and 'sph-lv' for short.

	i-vec	DNN i-vec	ph-vec	sph-lv
Male				
EER	5.527	4.695	5.177	4.192
DCF08	2.868	2.515	2.806	2.234
Female				
EER	7.070	5.477	6.537	5.407
DCF08	3.748	2.962	3.482	2.801

concatenated vectors of the selected local vectors φ_c and φ_l could be scored on the new PLDA model with the parameter set $\Theta = \{\tilde{\omega}, \tilde{\Psi}, \tilde{\Phi}, \tilde{\Lambda}\}$ where $\tilde{\omega}$, $\tilde{\Psi}$, $\tilde{\Phi}$ and $\tilde{\Lambda}$ are concatenated by the corresponding subcomponents from ω , Ψ , Φ and Λ .

5. Experiments

We focused on the text-constrained speaker verification task using the RSR2015 dataset [20]. The dataset consists of recordings from 300 speakers (143 female and 157 male) recorded in 9 sessions with multiple handphones and tablets. For both female and male speakers, all the utterances of speakers 1 to 25 were used to train the DNN; the utterances in part III of speakers 26 to 50 were used to train the variability models; the utterances in part III of the remaining speakers were used for system evaluation, where the lexicons are limited to the set of digits from 0 to 9. We constructed 1,753,850 trials (9,857 target and 861,488 non-target trials for male, 11,719 target and 870,786 non-target trials for female) where the 10-digit and 5-digit utterances were used for enrolment and test, respectively.

To derive the phone labels for the DNN training, we first train a GMM-HMM model to carry out force alignment. The GMM-HMM and the DNN of 183 output nodes were trained in a gender-independent manner. Since the lexicons are limited to 10 digits, containing 22 phones, the corresponding $22 \times 3 = 66$ output nodes were extracted and fine-tuned respectively for two gender-dependent DNNs, which were then used for the subsequent gender-dependent variability modeling.

For speaker modeling, 39-dimensional vectors of Mel frequency cepstral coefficients (MFCC) with first and second derivatives were used. The covariance matrices of the UBMs were diagonal. Three systems were compared:

- I. The GMM i-vector - 400-dimensional i-vectors were estimated on 64-Gaussian UBMs.
- II. The DNN i-vector - 66 Gaussian-like units were applied and 400-dimensional i-vectors were estimated.
- III. The phone-centric local variability vectors - The Gaussian-like units were the same with that used in II. The dimensionality of the latent variable was 20, resulting in a $20 \times 22 = 440$ variability vector via concatenation for every utterance.

For the PLDAs that followed, the covariance matrices were diagonal; and the speaker subspaces were of rank 200. To achieve the maximum modeling ability, the rank of the channel subspaces of the PLDA models were set to be the same with the lengths of the variability vectors, i.e., 400 for the GMM and DNN i-vectors while 440 for the phone-centric variability vector. The PLDA models were trained on the utterances drawn from speakers 1 to 50 in a gender-dependent manner.

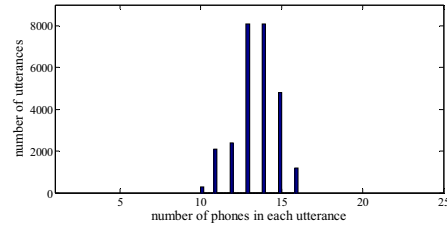


Figure 3: The number of the 5-digit utterances with respect to the number of phones contained in each utterance.

Table I reports the performances of the three systems in terms of EER and minDCF08 [21]. From the table, we observe that DNN i-vector and the phone-centric local variability vector outperform GMM i-vector. We also observe that the phone-centric variability vector is still inferior to DNN i-vector. This may be due to the content mismatch between the 10-digit enrollment and 5-digit test utterances. Fig. 3 shows the number of utterances where only 5 digits were included with respect to the number of phones. We can see that the 5-digit utterances contain in general 10 to 16 phones peaking around 13 or 14. For the 5-digit utterances, with a DNN i-vector shared by all the phones, the variability of the absent phones got approximated, whereas the phone-centric local vectors of the missing phones were ignorant of session variability, making the comparison between the 10-digit and 5-digit utterances unequal.

To cope with the content mismatch between the enrolment and test utterances, we scored the trials on the selected local vectors with the corresponding PLDA model parameters accordingly. In each trial, the local vectors of the enrolment and test utterance are selected according to the phones existing in the test utterance. Such selection can be done according to the phone recognition with an ASR system or with a pre-known text transcript. Then the variability vectors concatenated by the selected local vectors were subjected to the PLDA model which was composed of the subcomponents associated with the selected local vectors. We report the results in the last column of Table I, which show the best performance across all implementations. With this, we validate the idea that the flexibility of the phone-centric local vectors allows us to construct the vectors for speaker comparison based on the phonetic content of the utterances, providing a solution to content mismatch in text-constrained speaker comparison.

6. Conclusion

In this paper, we proposed the phone-centric local variability model for the text-constrained speaker verification task. The proposed model is implemented using the frame posterior probability derived with a deep neural network (DNN) trained for automatic speech recognition task. Experiments conducted on part III of RSR2015 dataset showed that the proposed phone-centric local variability vectors outperform the existing GMM and DNN i-vectors when the content of the test utterance is different from that of the enrolment. This result shows the potential of the proposed phone-centric local variability model in dealing with the content mismatch problem.

7. Acknowledgement

This work of Liping Chen is supported by the National Nature Science Foundation of China (Grant No. 61273264) and the electronic information industry development fund of China (Grant No. 2013-472).

8. References

- [1] D.A. Reynolds, T.F. Quatieri, and R.B. Dumm, "Speaker verification using adapted Gaussian mixture model," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-41, 2000.
- [2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12-40, Jan. 2010.
- [3] P. Kenny, G. Boulianne and P. Dumouchel, "Eigenvoice Modeling with Sparse Training Data," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 345-359, May 2005.
- [4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio Speech and Language Processing*, vol. 19, no. 4, pp. 788-798, May 2011.
- [5] P. Kenny, M. Mihoubi, and P. Dumouchel, "New MAP estimators for speaker recognition," in *Proc. The 8th European Conference on Speech Communication and Technology*, 2003, pp. 2691-2964.
- [6] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. International Conference on Computer Vision*, 2007.
- [7] A. Larcher, K. A. Lee, B. Ma, and et al, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60: 56-77, 2014.
- [8] T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann, and P. Dumouchel, "Text-Dependent Speaker Recognition using PLDA with Uncertainty Propagation," in *Proc. INTERSPEECH*, Aug. 2013, pp. 3684-3688.
- [9] N. Scheffer and Y. Lei, "Content matching for short duration speaker recognition," in *Proc. INTERSPEECH*, 2014, pp. 1317-1321.
- [10] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72-83, Jan. 1995.
- [11] L. Chen, K. A. Lee, B. Ma, W. Guo, H. Li and L. R. Dai, "Local variability vector for text-independent speaker verification," in *Proc. Odyssey: Speaker and Language Recognition Workshop*, Jun. 2014, pp. 54-59.
- [12] Y. Lei, N. Scheffer, L. Ferrer and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. ICASSP*, 2014, pp. 1714 - 1718.
- [13] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet and J. Alam, "Deep neural networks for extracting Baum-Welch statistics for speaker recognition," in *Proc. Odyssey: Speaker and Language Recognition Workshop*, Jun. 2014, pp. 293-298.
- [14] N. Scheffer, Y. Lei, "Content matching for short duration speaker recognition," in *Proc. INTERSPEECH*, Sep. 2014, pp. 1317-1321.
- [15] P. Kenny, "A small foot-print i-vector extractor," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2012, pp. 1- 6.
- [16] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proc. IEEE* vol. 77, no. 2, pp: 257-286, Feb. 1989.
- [17] G. E. Dahl and G. Hinton, "Acoustic modeling using deep Belief networks," *IEEE Trans. Audio Speech and Language Processing*, vol. 20, no. 1, pp.14-22, Jan. 2012.
- [18] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
- [19] S. Prince, *Computer vision: models, learning and inference*, Cambridge University Press, 2012.
- [20] Anthony Larcher, Kong-Aik Lee, Bin Ma, and Haizhou Li, "RSR2015: Database for text-dependent speaker verification using multiple pass-phrases.," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2012.
- [21] National Institute of Standards and Technology, *The NIST 2008 SRE Evaluation Plan*, 2008.