DEEP NEURAL NETWORKS BASED SPEAKER MODELING AT DIFFERENT LEVELS OF PHONETIC GRANULARITY

Yao Tian[1], *Liang He*[1], *Meng Cai*[2], *Wei-Qiang Zhang*[1], *Jia Liu*[1]

[1]National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing, China
[2]Microsoft Research Asia, Beijing, China

tianyao11@mails.tsinghua.edu.cn, heliang@mail.tsinghua.edu.cn
menca@microsoft.com, {zhangwq, liuj}@mail.tsinghua.edu.cn

ABSTRACT

Recently, a hybrid deep neural network/i-vector framework has been proved effective for speaker verification, where the DNN trained to predict tied-triphone states (senones) is used to produce frame alignments for sufficient statistics extraction. In this work, in order to better understand the impact of different phonetic precision to speaker verification tasks, three levels of phonetic granularity are evaluated when doing frame alignments, which are tied-triphone state, monophone state and monophone. And the distribution of the features associated to a given phonetic unit is further modeled with multiple Gaussians rather than a single Gaussian. We also propose a fast and efficient way to generate phonetic units of different granularity by tying DNN's outputs according to the clustering results based on DNN derived senone embeddings. Experiments are carried out on the NIST SRE 2008 female tasks. Results show that using DNNs with less precise phonetic units and more Gaussians per phonetic unit for speaker modeling generalize better to different speaker verification tasks.

Index Terms— speaker verification, deep neural networks, phonetic granularity

1. INTRODUCTION

Over recent years, many approaches based on Gaussian Mixture Model-Universal Background Model (GMM-UBM) have been proposed to improve the performance of speaker verification, among which i-vector has become a dominant approach in state-of-the-art speaker verification systems [1, 2, 3, 4]. In the i-vector paradigm, a sequence of acoustic feature vectors are mapped into a low-dimensional space and each utterance can be represented as a fixed-length vector called i-vector in this subspace. After the extraction of i-vectors, probabilistic linear discriminant analysis (PLDA) can then be applied as the backend classifier to get the final verification scores [5, 6, 7].

More recently, a hybrid framework which combines deep neural networks (DNN) in automatic speech recognition (ASR) with the conventional i-vector model has shown promising results for speaker verification [8, 9]. In their work, each output of the DNN is treated as a single Gaussian in the UBM and the posterior probability of the DNN output is then used as the occupation probability for sufficient statistics extraction [8]. Many later work based on similar concept of combining phonetically-related models have also been proved very effective for speaker verification [10, 11, 12].

In fact, senone is the basic phonetic unit for context-dependent triphone acoustic modeling in ASR and it performs much better than context-independent phonetic units [13, 14]. However, as for text-independent speaker verification, using such precise phonetic units for frame alignments might lead to severe phonetic content mismatch problem, especially on short-utterance and multilingual tasks. In addition, many previous works try to exploit phonetic information for speaker modeling by defining a set of phonetically-related GMMs rather than a single UBM [15, 16, 17, 18]. And it seems more necessary to investigate using GMMs for phonetic units modeling under the DNN/i-vector framework, where the DNN is discriminatively trained and the distribution of the features associated with each DNN's output might be much more complex.

Based on the above two problems, in this paper we analyze using DNNs with phonetic units of different granularity and using different number of Gaussian per phonetic unit for speaker modeling, trying to better understand the impact of different phonetic and acoustic precision to speaker verification tasks. Three levels of phonetic granularity are evaluated in this work which are tied-triphone state, monophone state and monophone, from context-dependent to context-independent. In addition, we also propose a fast and efficient way of generating phonetic units of different granularity by tying the DNN's outputs according to the clustering results based on senone embeddings. Here senone embedding is represented as the concatenation of DNN's weight vectors and biases. It should be noticed that the work in [19] has also explored the impact of different phonetic precision to speaker verification with senone tying. In their work, they represent each senone output as a single Gaussian and try to cluster senones by minimizing the likelihood loss of the UBM training data given the merge of similar Gaussians. Compared with their work, the senone tying method we propose in this paper is much more efficient and no additional training data are needed for senone clustering. We evaluate the proposed approaches on three kinds of tasks of NIST SRE 2008 female data set, which are English long utterance, multilingual long utterance, English short utterance. Experimental results show that systems using DNNs with less precise phonetic units and using more Gaussians to model the distribution of each phonetic unit generalize better to different kinds of speaker verification tasks.

The rest of this paper is organized as follows. Section 2 gives the DNN based i-vector framework. Section 3 presents the speaker modeling method at different levels of phonetic granularity. Section 4 introduces the senone tying method based on senone embeddings. Experimental setup and results are given in Section 5. Conclusions are presented in Section 6.

2. THE DNN BASED I-VECTOR FRAMEWORK

Given a speech segment, the following sufficient statistics (Baum-Welch statistics) which are used for i-vector subspace training and i-vector extraction need to be calculated using UBM

$$N_c = \sum_t p(c|\mathbf{x}_t, \lambda) \quad (1)$$

$$\mathbf{F}_c = \sum_t p(c|\mathbf{x}_t, \lambda)\mathbf{x}_t \quad (2)$$

$$\mathbf{S}_c = \sum_t p(c|\mathbf{x}_t, \lambda)\mathbf{x}_t\mathbf{x}_t^T \quad (3)$$

where N_c, $\mathbf{F}_c$ and $\mathbf{S}_c$ are the zero-order, first-order and second-order statistics with respect to c-th Gaussian. λ represents the UBM model and $\lambda_c = \{\pi_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}, c = 1, 2, ..., C$. $\mathbf{x}_t$ is the acoustic feature. $p(c|\mathbf{x}_t, \lambda)$ is the alignment of $\mathbf{x}_t$ calculated as the posteriors of the c-th Gaussian component.

By treating each senone (tied-triphone state) output of the DNN as a single Gaussian in the UBM, Lei [8] and Kenny [9] propose to use the senone posteriors generated by the DNN to do the frame alignments as a replacement of the UBM

$$p(c|\mathbf{x}_t, \lambda) \leftarrow p(s|\mathbf{x}_t, \tau) \quad (4)$$

where τ represents the DNN model. s is the index of the DNN's output nodes which correspond to senones. In this way, the text-independent speaker verification task has been converted as a phonetically-dependent (senone-dependent) one.

3. SPEAKER MODELING AT DIFFERENT LEVELS OF PHONETIC GRANULARITY

The single Gaussian assumption in the DNN/i-vector framework might not be so appropriate since the DNN is discriminatively trained and the distribution of the features associated with each DNN's output could be much more complex, not to mention that the features for frame alignment and statistics calculation are usually completely different [8]. In order to better describe the acoustic space, we use mutiple Gaussians rather than a single Gaussian to represent each phonetic unit's distribution when building the DNN/i-vector system. The GMM corresponding to each phonetic unit could be estimated as follows in an iterative way as in the traditional EM algorithm

$$\gamma_{c,s,t} = p(c|\mathbf{x}_t, \lambda_s)p(s|\mathbf{x}_t, \tau) \quad (5)$$

$$N_{c,s} = \sum_t \gamma_{c,s,t} \quad (6)$$

$$\pi_{c,s} = \frac{N_{c,s}}{\sum_c N_{c,s}} \quad (7)$$

$$\boldsymbol{\mu}_{c,s} = \frac{\sum_t \gamma_{c,s,t}\mathbf{x}_t}{N_{c,s}} \quad (8)$$

$$\boldsymbol{\Sigma}_{c,s} = \frac{\sum_t \gamma_{c,s,t}\mathbf{x}_t\mathbf{x}_t^T}{N_{c,s}} - \boldsymbol{\mu}_{c,s}\boldsymbol{\mu}_{c,s}^T \quad (9)$$

where $\gamma_{c,s,t}$ is the posterior probability corresponding to the c-th Gaussian of phonetic unit s given acoustic feature $\mathbf{x}_t$. Now each phonetic unit is represented by a GMM model λ_s and $\lambda_{c,s} = \{\pi_{c,s}, \boldsymbol{\mu}_{c,s}, \boldsymbol{\Sigma}_{c,s}\}$.

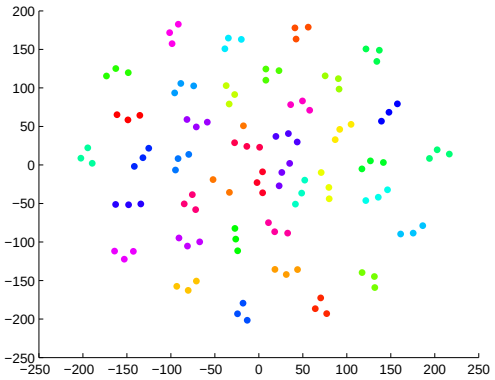

Fig. 1. *The visualization of monophone states embeddings using t-SNE toolkit. The nodes of the same color belong to the same phone.*

As for i-vector modeling, we first extract sufficient statistics based on each phonetic unit's GMM. Then these sufficient statistics are concatenated for i-vector subspace training and i-vector extraction. And the posterior probability corresponding to each Gaussian is modified as

$$p(c, s|\mathbf{x}_t) \leftarrow p(c|\mathbf{x}_t, \lambda_s)p(s|\mathbf{x}_t, \tau) \quad (10)$$

It should be noticed that this is a soft alignment process. In fact, some other works [16, 18] have also investigated using GMM to model each phonetic unit for speaker verification but they use a hard alignment and usually they use an ASR decoding system to produce frame alignments.

4. SENONE TYING WITH DNN DERIVED EMBEDDINGS

In fact, different GMM-HMM systems have to be firstly trained to generate the transcriptions of each level's phonetic units and this would cost too much time. In this paper, we also propose a fast and efficient senone tying method based on DNN derived senone embeddings to generate phonetic units of different granularity.

The DNN can actually be seen as a joint model combing feature learning and a log-linear classifier. The raw acoustic feature is firstly converted into a highly discriminative feature representation through the many hidden layers of nonlinear transforms. Then a log-linear model (output layer) is used to generate the classification results based on the feature from the last hidden layer.

Let's denote the weight matrix and the bias of the output layer as $\mathbf{W}_{H\times S}$ and $\mathbf{b}_S$, where H denotes the number of hidden units in the last hidden layer and S denotes the number of output senones. Each senone corresponds to a specific column of $\mathbf{W}_{H\times S}$ and an element of $\mathbf{b}_S$. They are concatenated to form a vector called senone embedding in this paper. Actually, if two senones have similar posteriors given the same features, the senone embeddings would be similar as well. So it's reasonable to use senone embeddings to measure the similarity of different senones. The senones are clustered in a bottom-up way based on the Euclidean distance measure where two senones with the minimum distance are put together in each iteration. And the average of the senone embeddings that have been grouped to the same cluster is used to represent the new senone for next iteration. After the clustering, the summation

of the outputs' values in the same cluster are regarded as the posterior probability of the new phonetic unit for sufficient statistics extraction.

We try to verify the effectiveness of the proposed methods by visualizing monophone states based embeddings where each monophone is modeled with three states. The phone set is from CMU dictionary [20] and contains 39 phones. The visualization is based on the t-Distributed Stochastic Neighbor Embedding (t-SNE) toolkit [21] which uses t-SNE model for dimensionality reduction and is particularly well suited for the visualization of high-dimensional datasets. All the embeddings are reduced to two-dimensional vectors based on t-SNE toolkit and are plotted in Figure 1. We can see that embeddings belong the the same phone are very close to each other.

5. EXPERIMENTS

5.1. Experimental setup

5.1.1. Dataset

Experiments are carried out on three types of speaker verification tasks of NIST SRE 2008 female telephone data, which are English long utterance (short2-short3 condition7), multilingual long utterance (short2-short3 condition6), English short utterance (short2-10sec, 10sec-10sec). The duration of long (short2, short3) and short (10sec) utterances are approximately 2.5 minute and 10 second respectively. The training data for phonetically-related DNNs are 300 hours English telephone speeches from Switchboard-I. The training data for UBM, i-vector $\mathbf{T}$ matrix and PLDA are selected from NIST SRE 04, 05, 06 telephone data.

5.1.2. Models

- **GMM-HMM**: A GMM-HMM system based on HTK toolkit [22] is firstly trained to generate the transcriptions of each level's phonetic unit. The CMU dictionary is used which contains 39 phones [20]. 52-dimensional PLP features (13 basic + first/second/third order) are used with speaker-based mean-covariance normalization. Then the features are reduced to 39 dimension by HLDA. The transcriptions of monophone, monophone state based systems are generated from a monophone GMM-HMM model, while the transcriptions of tied-triphone state based system are generated from a triphone GMM-HMM model. The number of tied-triphone states is decided by a phonetic decision tree.
- **DNN**: All the DNNs used in this paper have five hidden layers and are fine-tuned with cross-entropy criterion based on the transcriptions generated by GMM-HMM systems. The input of the DNN is a concatenation of 11 frames and each frame consists of 120 log Mel-filterbank coefficients (40 basic + first/second order). Each hidden layer has 1200 nodes. The number of output's nodes are either 2227, or 1038, or 521, or 117 for tied-triphone state, 117 (39 phones $\times$ 3 states) for monophone state, 39 (39 phones) for monophone.
- **UBM/i-vector model**: The acoustic feature adopted is 39-dimensional (13 basic + first/second order) PLP feature. Then a gender-dependent diagonal covariance UBM with 2048 mixtures is trained. The dimensionality of i-vectors is 400. Simplified Gaussian PLDA [7] is used to generate scores and the dimensionality of speaker subspace in PLDA model is 200.
- **DNN/i-vector model**: The DNN is used to provide frame posteriors. Then they are combined with 39-dimensional PLP features for sufficient statistics extraction. The total number of mixtures is confined by the number of phonetic units times the number of mixtures per phonetic unit. Other model configurations are the same with the UBM/i-vector model.

Equal error rate (EER) and minimum decision cost function (minDCF) are adopted for evaluation [23].

5.2. Experimental results

5.2.1. Speaker modeling at different levels of phonetic granularity

Experimental results of the baseline 2048 Gaussians UBM/i-vector system and the 2227 senones DNN/i-vector system are presented in the first two rows of Table 1. Results show that DNN/i-vector outperforms UBM/i-vector on short2-short3 condition7 (English long utterance) task. However, the performance of DNN/i-vector degrades sharply compared to the UBM/i-vector on short2-short3 condition6 (Multilingual long utterance) task. Actually more than ten different languages are involved in this task. DNNs trained with English phonetic units only can hardly describe the pronunciation patterns of so many languages. Besides, the tied-triphone state is a very precise phonetic unit and highly language-dependent which makes it even less suitable for multilingual task. It is interesting to see that DNN/i-vector is inferior to UBM/i-vector on short2-10sec and 10sec-10sec (English short utterance) tasks. We think phonetic content mismatch might be a reasonable explanation for the performance degradation. Since we are dealing with text-independent short utterance task, the frames of enrollment and test utterances are more likely to be aligned to different senones using discriminatively trained DNN and the comparison between them will be less effective. However, the problem might not be so severe in UBM based system where the generatively trained model usually has more blurred boundaries between different phonetic units.

The results of DNN/i-vector at different levels of phonetic granularity are presented in Table 1 as well. The number of Gaussians of the systems are all around 2000.

On short2-short3 condition7 (English long utterance) task, it can be seen that the best result is obtained with 521 senones based DNN/i-vector system, and the relative improvements are 17.7% in EER and 20.2% in minDCF compared with 2227 senones based approach. The performance becomes worse when the number of senones is further reduced. 117 is the minimum senone number we could obtained based on the HTK toolkit [22] and from the results we can see that the performance of 117 senones based system is similar to 117 monophone states based system since the triphone states that belong to the same phone are already clustered to the same senone. The 39 monophones based system performs slightly worse than 117 monophone states based approach. However, it still outperforms the 2227 senones based system. The above results actually indicate that the frame alignment process does not need to be so accurate on content. A balance of phonetic and acoustic precision brings more benefits to speaker verification.

On short2-short3 condition6 (multilingual long utterance) task, results show that consistent performance improvements could be obtained when we reduce the number of phonetic units. Compared with 2227 senones based system, the relative improvements of 39 monophones based approach are 23.8% in EER and 17.6% in minDCF. As the number of phonetic units becomes less, the phonetic unit becomes more general which makes it more suitable for multilingual tasks. It should be noticed the improvements are not so obvious compared with UBM based system since the DNN is still trained with English data only which makes the phonetic

Table 1. *Performance (EER(%)/minDCF08×10) of UBM/i-vector and DNN/i-vector at different levels of phonetic granularity on the NIST SRE08 female short2-short3 condition7, short2-short3 condition6, short2-10sec, 10sec-10sec tasks. (G denotes Gaussian)*

acoustic model	model size	short2-short3 condition7	short2-short3 condition6	short2-10sec	10sec-10sec
UBM	2048×1G	2.02/0.106	5.08/0.268	10.33/0.497	16.60/0.752
tied-triphone state	2227×1G	1.81/0.089	6.64/0.323	11.11/0.574	17.85/0.826
tied-triphone state	1038×2G	1.67/0.078	6.28/0.304	10.78/0.557	17.52/0.808
	521×4G	**1.49/0.071**	5.65/0.286	10.14/0.542	16.71/0.801
	117×16G	1.56/0.080	5.26/0.269	9.75/0.508	16.11/0.748
monophone state	117×16G	1.58/0.081	5.11/0.271	9.72/0.495	16.12/0.742
monophone	39×48G	1.65/0.082	**5.06/0.266**	**9.66/0.472**	**15.87/0.735**

Table 2. *Performance (EER(%)/minDCF08×10) of UBM/i-vector and DNN/i-vector with DNN's senone outputs tying on the NIST SRE08 female short2-short3 condition7, short2-short3 condition6, short2-10sec, 10sec-10sec tasks. (G denotes Gaussian)*

acoustic model	model size	short2-short3 condition7	short2-short3 condition6	short2-10sec	10sec-10sec
UBM	2048×1G	2.02/0.106	5.08/0.268	10.33/0.497	16.60/0.752
tied-triphone state-2227	2227×1G	1.81/0.089	6.64/0.323	11.11/0.574	17.85/0.826
	1038×2G	1.72/0.088	6.43/0.326	10.95/0.575	17.63/0.821
	521×4G	**1.54/0.076**	5.74/0.294	10.21/0.562	16.61/0.796
	117×16G	1.62/0.082	5.35/0.278	9.81/0.513	16.23/0.764
	39×48G	1.65/0.084	**5.19/0.275**	**9.70/0.481**	**15.95/0.741**
tied-triphone state-8223	2227×1G	1.83/0.091	6.68/0.330	11.12/0.577	17.88/0.822
	521×4G	1.56/0.078	5.77/0.299	10.18/0.568	16.62/0.792
	39×48G	1.68/0.085	5.21/0.282	9.75/0.485	16.01/0.743

units are language dependent. UBM trained without phoneme-specific tuning might still be an ideal option for different languages since it has a more blurred clustering which might somehow benefit multilingual tasks. We will further investigate using DNNs trained with multilingual data in our later work.

Results on short2-10sec and 10sec-10sec (English short utterance) show that systems' performance could be improved when we reduce the number of phonetic units. Compared with 2227 senones based system, the relative improvements are 13.1% in EER and 17.8% in minDCF on short2-10sec task, 11.1% EER and 11.0% in minDCF on 10sec-10sec with 39 monophones based system. When we reduce the number of phonetic units, the frames of enrollment and test utterances have greater chance to be aligned to the same phonetic units first and the further frame alignments based on GMMs might be more balanced which leads to less severe phonetic content mismatch problem. This might be the reason that the performance becomes better when the phonetic unit is not so precise.

5.2.2. Speaker modeling based on unsupervised senone tying

The results of DNN/i-vector systems based on unsupervised senone tying are presented in Table 2. We first conduct the experiments on DNN with 2227 senones. From the results in Table 2 we can see that competitive results could be obtained compared with supervised decision tree based senone tying which demonstrate the effectiveness of the proposed senone tying method.

In fact, the DNN model used in speaker verification is trained based on the same criteria as in speech recognition. However, the senone number is usually much larger (around 8000 or 9000) for speech recognition and we have to train a DNN with less senones for speaker verification separately. As a result, we'll have to go through the DNN forward process twice if we want to do speech recognition and speaker verification tasks at the same time. So we also experiment the senone tying method on DNN with 8223 senones to see if it is feasible for speaker verification using DNN with larger senones. From the results in Table 2 we can see the method is effective and similar performance could be obtained. These results indicate that it is practical to use the same DNN for speech recognition and speaker verification while maintaining good performance for both tasks.

6. CONCLUSIONS

In this paper we assess the DNN/i-vector approach for speaker verification by using DNNs with phonetic units of different granularity. Experimental results show that a balance of phonetic and acoustic precision brings more benefits to speaker verification and generalize better in the presence of phonetic mismatch. In addition, the proposed senone tying method based on senone embeddings is effective for speaker verification and it demonstrates that it is practical to use DNNs with larger number of senones for speaker verification when senone tying is used. In the future, we'll further investigate using multilingual DNNs for speaker modeling.

7. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China under Grant No. 61273268, No. 61370034 and No. 61403224.

8. REFERENCES

[1] 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.

[2] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[3] 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, 2010.

[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[5] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE 11th International Conference on Computer Vision, 2007*. IEEE, 2007, pp. 1–8.

[6] P. Kenny, "Bayesian speaker verification with heavy-tailed priors." in *Odyssey*, 2010, p. 14.

[7] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems." in *Interspeech*, 2011, pp. 249–252.

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

[9] P. Kenny, T. Stafylakis, P. Ouellet, V. Gupta, and J. Alam, "Deep neural networks for extracting Baum-Welch statistics for speaker recognition," in *Odyssey 2014*, 2014, pp. 293–298.

[10] Y. Tian, M. Cai, L. He, and J. Liu, "Investigation of bottleneck features and multilingual deep neural networks for speaker verification." in *Interspeech*, 2015, pp. 1151–1155.

[11] P. Mat, J. H. Cernock *et al.*, "Analysis of DNN approaches to speaker identification." in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5100–5104.

[12] Y. Tian, M. Cai, L. He, Z. Wei-Qiang, and J. Liu, "Improving deep neural networks based speaker verification using unlabeled data." in *Interspeech*, 2016, pp. 1863–1867.

[13] A.-r. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[14] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[15] A. Park and T. J. Hazen, "ASR dependent techniques for speaker identification," in *ICSLP 2012*, 2002, pp. 1337–1340.

[16] B. J. Baker, R. J. Vogt, and S. Sridharan, "Gaussian mixture modelling of broad phonetic and syllabic events for text-independent speaker verification," in *Eurospeech 2005*, 2005, pp. 2429–2432.

[17] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Compensation of nuisance factors for speaker and language recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 15, no. 7, pp. 1969–1978, 2007.

[18] Z.-Y. Li, W.-Q. Zhang, W.-W. Liu, Y. Tian, and J. Liu, "Text-independent speaker verification via state alignment," in *Odyssey*, 2014, pp. 68–72.

[19] S. Cumani, P. Laface, and F. Kulsoom, "Speaker recognition by means of acoustic and phonetically informed GMMs," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[20] "The CMU pronouncing dictionary," *http://www.speech.cs.cmu.edu/cgi-bin/cmudict/*.

[21] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[22] 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," *Cambridge university engineering department*, vol. 3, p. 175, 2002.

[23] "The NIST year 2008 speaker recognition evaluation plan," *http://www.itl.nist.gov/iad/mig//tests/sre/2008/*, 2008.