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

DNN i-vector Speaker Verification with Short, Text-constrained Test Utterances

Jinghua Zhong[1], *Wenping Hu*[2], *Frank Soong*[2], *Helen Meng*[1]

[1]Department of Systems Engineering & Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR of China,
[2]Speech Group, Microsoft Research Asia, Beijing, China
[1]{jhzhong, hmmeng}@se.cuhk.edu.hk, [2]{wenh, frankkps}@microsoft.com

Abstract

We investigate how to improve the performance of DNN i-vector based speaker verification for short, text-constrained test utterances, e.g. connected digit strings. A text-constrained verification, due to its smaller, limited vocabulary, can deliver better performance than a text-independent one for a short utterance. We study the problem with "phonetically aware" Deep Neural Net (DNN) in its capability on "stochastic phonetic-alignment" in constructing supervectors and estimating the corresponding i-vectors with two speech databases: a large vocabulary, conversational, speaker independent database (Fisher) and a small vocabulary, continuous digit database (RSR2015 Part III). The phonetic alignment efficiency and resultant speaker verification performance are compared with differently sized senone sets which can characterize the phonetic pronunciations of utterances in the two databases. Performance on RSR2015 Part III evaluation shows a relative improvement of EER, i.e., 7.89% for male speakers and 3.54% for female speakers with only digit related senones. The DNN bottleneck features were also studied to investigate their capability of extracting phonetic sensitive information which is useful for text-independent or text-constrained speaker verifications. We found that by tandeming MFCC with bottleneck features, EERs can be further reduced.
Index Terms: DNN i-vector, DNN adaptation, senone, frame alignment

1. Introduction

Speaker verification is the process to accept or reject a person's identify claim through his/her voice. It usually falls into two types: text-independent [1] and text-dependent [2]. Text-independent speaker verification (TISV) doesn't have any constraint on the text content. People have the freedom to say whatever they want to. It has been successfully applied to military intelligence and forensic tasks. But large amounts of development data and long utterances needed in text-independent domain is not applicable for commercial application. Text-dependent speaker verification (TDSV) requires the speaker to utter certain pass-phases during the authentication. The matched content makes speaker verification with short utterances possible. It is not very flexible from the user's point of view. Besides, it is possible that an imposter can record utterance from a user beforehand and then play it back. The text-constrained speaker verification can avoid these issues to a certain extent by only restricting the vocabulary instead of fixed phrase. Digit is the most common used fixed vocabulary. If it is generated randomly and prompted to the user during verification, it is claimed that it becomes harder for anyone to break in instead of you. Besides, due to the smaller and limited vocabulary in text-constrained speaker verification, it can also deliver better performance than a text-independent one for a short utterance. In this work, we focus on speaker verification using randomly prompted digit strings.

In the past decades, most researches have been focused on the more challenging text-independent speaker verification. In this field, Gaussian Mixture Model (GMM)[3] based i-vector [4] has become a popular approach for many systems. It compresses both channel and speaker information into a low-dimensional space called total variability space, and accordingly projects each GMM supervector to a total factor feature vector called the i-vector. Then Linear Discriminant Analysis (LDA)[5] and Probabilistic LDA (PLDA)[6] were applied on the i-vectors for inter-session compensation. In [7], a deep neural network (DNN) trained for automatic speech recognition (ASR) was used to substitute the role of GMM. It used DNN senone posterior as frame alignment in i-vector extraction process. The phonetic information provided through senone posteriors can improve the accuracy of frame alignment and therefore achieve better speaker verification performance.

The above i-vector approaches have been proven to be very effective for text-independent speaker verification based on long utterances. In [8], the i-vector extractor was trained on text-independent NIST data and the effect of adding phonetic information to speaker classes into PLDA training for text-dependent speaker verification were assessed. In [9, 10], Stafylakis et al. proposed to train the i-vector extractor on short utterances directly and then dealt with the phonetic variability in a PLDA classifier by making the PLDA model parameters phrase-dependent. These attempts of using i-vectors led to disappointing results because the i-vector representation of short utterances is sensitive to the phonetic content. For speaker verification with random digit strings, Kenny et al. [11] proposed several different ways of using Joint Factor Analysis (JFA) as a feature extractor and a Joint Density model as a backend to estimate likelihood ratios. The use of speech recognition method of force alignment to extract phonetically-aware Baum-Welch statistics achieved better results on RSR2015 Part III [12] evaluation. In [13, 14], Chen et al. proposed the phone-centric local variability model for the content-matching at the statistics level.

In this work, we study the problem of speaker verification with random digit strings on RSR2015 Part III evaluation. We first investigate phonetically aware DNN in its capability on estimating supervectors and the corresponding i-vectors with two speech databases: Fisher and RSR2015 Part III. Through both Universal Background Model (UBM) component posteriors and DNN senone posteriors, most frames are aligned to only

The work was done during the first authors internship in the Speech Group of Microsoft Research Asia.

http://dx.doi.org/10.21437/Interspeech.2017-1036

a few Gaussian components or senones, especially for short utterances with fixed vocabulary. Statistics estimation of those components with insufficient frames are biased [15]. In this case, we use differently sized senone sets to characterize the phonetic pronunciations of utterances in the two databases. The phonetic alignment efficiency and resultant speaker verification performance are compared between the two senone sets. Except for providing posteriors in statistics estimation, DNNs could also be used to extract phonetic discriminant features with a bottleneck hidden layer. The DNN bottleneck feature can complement speaker-dependent phonetically discriminative information for the acoustic features. Besides, MFCC feature could not reflect some speaker characteristics associated with the high frequency range of speech which is down-sampled by the mel scale [16]. In this work, we try to investigate whether bottleneck features could also make up these speaker information.

The rest of the paper is organized as follows. In Section 2, we describe the background of DNN i-vector and DNN bottleneck feature. Then we describe how we build the DNN i-vector systems for text-constrained speaker verification in Section 3. Implementation and experimental results on the RSR2015 Part III corpus are presented in Sections 4. Finally, the conclusions are presented in Section 5.

2. Background

2.1. DNN i-vector

The i-vector approach is based on JFA. Channel factors in JFA are supposed to model only channel effects, also contain information about speakers. In [4], Dehak et al. proposed to define a new low-dimensional space to model both speaker and channel variabilities. Given features of N utterances and N_u frames for the u-th utterance, $\{x_i^{(u)}\}_{\{i=1,\ldots,N_u;u=1,\ldots,N\}}$, F is the dimension of each frame, the i-th speech frame $x_i^{(u)}$ from the the u-th utterance is assumed to be generated by the following Gaussian distribution:

$$x_i^{(u)} \sim \sum_k \pi_k^{(u)} \mathcal{N}(m_k + T_k \omega^{(u)}, \Sigma_k) \tag{1}$$

where T_k matrices describe a low-rank space (named total variability space) and $\omega^{(i)}$ is a low-dimensional total variability factor (named i-vector) with standard normal distribution. In the baseline GMM i-vector approach, m_k and Σ_k is the mean and covariance of the k-th Gaussian in UBM. There are K Gaussian components in the UBM which is used as the class k in Eq. 1. Here, the frame alignments of $x_i^{(u)}$ are done by the posterior of the k-th Gaussian $\gamma_{ki}^{(u)}$ in UBM.

In [7], Lei et al. proposed to use DNN trained for ASR to substitute GMM in the i-vector extraction process. In the state-of-the-art ASR systems, the pronunciations of all words are represented by a sequence of senones Q. Each senone, determined by a decision tree using the maximum likelihood (ML) approach, is used to model the triphone states.

Lei et al. [7] proposed to use the senones as the classes k in Eq. 1, instead of the Gaussian indices in the GMM i-vector. Then a DNN is trained to predict the posteriors $\gamma_{ki}^{(u)}$ for each of the k classes, defined as senones now, as the frame alignments for $x_i^{(u)}$. Given a speech utterance, the Baum-Welch statistics can be computed using the posterior probabilities of the senone classes:

$$\begin{aligned} N_k^{(u)} &= \sum_i \gamma_{ki}^{(u)} \\ F_k^{(u)} &= \sum_i \gamma_{ki}^{(u)} (x_i^{(u)}) \end{aligned} \tag{2}$$

These sufficient statistics are used to train the total variability matrix T and extract the i-vector $\omega^{(i)}$. We can also get the means m_k and covariance Σ_k of the senones defined in Eq. 1:

$$\begin{aligned} m_k &= \frac{\sum_{i,u} \gamma_{ki}^{(u)} x_i^{(u)}}{\sum_{i,u} \gamma_{ki}^{(u)}} \\ \Sigma_k &= \frac{\sum_{i,u} \gamma_{ki}^{(u)} x_i^{(u)} x_i^{(u)^T}}{\sum_{i,u} \gamma_{ki}^{(u)}} - m_k m_k^T \end{aligned} \tag{3}$$

DNN can provide phonetic information while GMM is "phonetically unaware". So DNN senone posteriors can improve the accuracy of frame alignment.

2.2. DNN bottleneck feature

Except for providing posteriors in statistics estimation, DNN can also be used as a means of feature extraction. One of the hidden layers has a small number of nodes relative to the size of the other hidden layers. This hidden layer, named bottleneck layer, uses a linear activation and the activation is used as a feature vector, named bottleneck feature [17]. With the acoustic feature as input and phonetic senone as output, the bottleneck features extracted from the bottleneck DNN contain speaker-dependent phonetically discriminative information [18]. Since the bottleneck features should have little speaker information, we concatenate the bottleneck feature and acoustic feature as tandem feature to compute the sufficient statistics.

3. Building DNN i-vector systems for Text-constrained Speaker Verification

3.1. DNN adaptation and digit senone selection

DNN i-vector was proposed to replace GMM-UBM i-vector by incorporating senone posteriors in constructing better phonetically aligned supervectors. However, using a large number of DNN outputs for ASR, usually in the thousands, it needs a large amount of transcribed data. The RSR background and development data, only about 23h, are too small to train a DNN for ASR. So we train the ASR DNN with the Fisher corpus. 3504 senones were defined for the Fisher corpus determined by a decision tree. We leverage insufficient in-domain transcribed RSR data by DNN adaptation. With the adapted DNN model, we realign the RSR data and get more accurate alignments for DNN adaptation. The adapted DNN model is used to compute senone posteriors for frame alignment.

The most difficult part of i-vector based approaches in short utterances seems to be the content mismatch between the enrollment and test utterances. Through both UBM component posteriors and DNN senone posteriors, most frames are aligned to only a few Gaussian components or senones, especially for short utterances with fixed vocabulary. The estimated posterior vectors tend to be sparse. Statistics estimations of those components with insufficient frames are biased. The senone set defined for the Fisher corpus needs to be large so that it could cover most common vocabulary. However, because of the vocabulary constraint of the English digit in RSR2015 Part III, the

corpus contains only a small part of this senone set. We proposed to only select the digit related senones for statistics estimation. So we first use the Fisher trained DNN-HMM model to do force alignment on RSR background and development set. We then get the digit senone set from the alignments. There are only 305 digit senones out of a total 3504 senones. After eliminating 3 senones associated with silence, 302 valid digit senones are used for i-vector extractor training. The Baum-Welch statistics are extracted for only 302 digit related senones. Before statistics estimation, we make the sum of posteriors to 302 selected senones equal to 1 for posterior normalization. In this case, we can get more accurate "stochastic phonetic-alignment" for constructing supervectors and estimating the corresponding i-vectors. The performance comparison of different size senone sets for characterizing the phonetic pronunciations of utterances are reported in section 4.2.

3.2. Bottleneck features vs. MFCC features

Bottleneck features have been verified by many previous works to complement phonetic information for the acoustic feature. So we investigate the capability of DNN bottleneck features in extracting phonetic sensitive information on text-constrained speaker verification. We use the bottleneck DNN trained on Fisher corpus to extract bottleneck feature.

On the other hand, MFCC feature has been dominantly used in both speech recognition and speaker recognition. But speech recognition extract phonetic information from speeech while speaker recognition extract speaker information from speech. MFCC was first proposed to mimic how human ears process sound for speech recognition. The mel scale in frequency makes the spectral resolution lower when the frequency increases. This would down-sample the spectral characteristics in the high frequency region. So that MFCC feature would not reflect some speaker characteristics associated with the high frequency range of speech [16]. However, speaker characteristics associated with the vocal tract length, are reflected more in the high frequency region of speech based on the theories in speech production [19]. The relatively shorter vocal tract in females makes higher formant frequencies in their speech. This maybe the reason why speaker recognition of female task is usually tougher than that of male task with MFCC feature. In this work, we try to investigate whether bottleneck feature from speech recognition could also make up the shortage of MFCC feature in the high frequency region.

4. Experimental results

4.1. Experimental setup

In the RSR2015 Part III corpus [12], each speaker enrollment contains three ten-digit sequences and each test utterance contains a five-digit sequence. The total duration is about 15s for enrollment and 3s for test. The entire Part III of the RSR2015 database consists of 34h and 36min of audio recording (12h and 51min of nominal speech after VAD). We used the RSR background and development sets for model training and the evaluation set for testing. The training set includes 100 male and 94 female speakers together with 22,661 utterances. For the male task, there are 57 target speakers and 3,419 test utterances, composed of the trial list with 3,419 true target trials and 191,464 imposter trials. For the female task, there are 49 target speakers and 2,938 test utterances, composed of the trial list with 2,938 true target trials and 141,024 imposter trials. Session non-overlap between enrollment and testing is maintained to maximize the mismatch.

In DNN i-vector system, both the HMM-GMM and HMM-DNN ASR models were trained on about 300 hours of clean English telephone speech from Fisher data set. The cross-word tri-phone HMM-GMM ASR with 3504 senones was trained with 39-dimensional MFCC features, including 13 static features and first and second order derivatives. A six-layer DNN with 585 input nodes, 2048 nodes in each hidden layer and 3504 output nodes was trained using the alignments from the HMM-GMM. The input layer of the DNN was composed of 15 frames (7-1-7) of 39-dimensional MFCC feature. The features were pre-processed with utterance-based MVN algorithm. DNN adaptation was done with all the background and development set from RSR2015 Part III. After eliminating 3 senones associated wih silence, 3501 valid senones were used for i-vector extractor training. The acoustic features for speaker modeling were the first 19 Mel frequency cepstral coefficients and log energy, together with their first and second derivatives. Energy-based voice-activity detection (VAD) and utterance-based cepstral mean and variance normalization (CMVN) were applied. These 60-dimensional feature vectors were used in the DNN system to compute sufficient statistics for a 400-dimensional i-vector extractor. The dimensionality of the i-vectors was further reduced by gender-independent LDA, followed by length normalization and gender-independent PLDA.

The bottleneck DNN was trained on same $\sim$300 hours Fisher data set. A six-layer bottleneck DNN with 585 input nodes, 2048 nodes in each hidden layer except 40 nodes bottleneck layer, and 3504 output nodes was trained, with the third hidden layer as bottleneck layer. The 40-dimensional bottleneck feature was concatenated with 20-dimensional MFCC feature, including the first 19 Mel frequency cepstral coefficients and log energy, to form 60-dimensional tandem feature.

4.2. DNN adaptation and digit senone selection

This set of experiments are to investigate the effectiveness of DNN adaptation and digit senone selection in DNN i-vector approach. The original DNN model trained with Fisher data (Fisher trained DNN) is regarded as baseline for comparison. The back-end of the Fisher trained DNN is based on all 3501 valid senones in i-vector extraction process. The RSR adapted DNN is adapted from the Fisher trained DNN using realignments of RSR training set. Here, we use all 3501 valid senones or only 302 digit related senones in i-vector extraction process for performance comparison. Table 1 summarize the results obtained with different DNN model and different senone sets on RSR2015 Part III male and female task.

Table 1: *DNN i-vector results on RSR2015 Part III evaluation. The notation in EER means male / female.*

Model	Used Senone No.	EER (%)
Fisher trained DNN	3501	1.90 / 2.54
RSR adapted DNN	3501	1.86 / 3.13
	302	**1.75 / 2.45**

From the table, we observe that: (1) Comparing the results of two DNN models using all 3501 valid senones, inconsistent performance improvement of only using DNN adaptation with RSR training data. A possible reason may be that the i-vectors extracted using the RSR adapted DNN are noisy, because almost all frames are aligned to digit related senones. Statistics estimations of other senones with insufficient frames are biased. (2) Comparing the results of RSR adapted DNN model using

different senone sets, using only 302 digit related senones for statistics estimation can significantly improve the performance especially for the female task (from 3.13% to 2.45%). Besides, it is more efficient with a small model size of i-vector extractor compressed from 3501×60 to 302×60.

We also evaluate the performance of the DNN trained for ASR with only about 23h RSR training data (RSR trained DNN) for comparison. We use the DNN trained on the Fisher corpus to do force alignment for RSR training data. A DNN with 585-nodes input layer, four hidden layers and 305-nodes output layer was trained using these alignments. The 305 nodes in output layer are defined as the previous selected digit related senone set. We investigate the effectiveness of DNN trained on insufficient in-domain training data with regards to the different number of nodes in hidden layers. The four hidden layers are with same number of nodes. From the results in Table 2, we can observe that: (1) For an authentication task with insufficient training data, we train a more robust DNN with a small model size. (2) When we train the DNN with 512 hidden nodes, the RSR trained DNN could obtain better results than RSR adapted DNN for male task but obtain worse results for female results.

Table 2: *Performance of RSR trained DNN i-vector approach with different number of hidden nodes on RSR2015 Part III evaluation. The notation in EER means male / female.*

Hidden nodes	EER (%)
2048	1.93 / 3.10
1024	1.84 / 3.03
512	**1.70 / 2.69**

Then we compare the three different DNN models, Fisher trained DNN, RSR adapted DNN and RSR trained DNN, on DET curve in Fig. 1. The three solid lines represents results of female task and the three dotted lines represents the results of male task. The lines with same color represent results of the same system. From the figure, we can see that the DET curves from three different DNN models are close to each other especially for the female task. So DNN seems to be robust for data set mismatch which is good for authentication tasks with insufficient in-domain training data.

4.3. Bottleneck features vs. MFCC features

The bottleneck DNN model is trained with Fisher data. The tandem feature consists of 20-dimensional MFCC feature and 40-dimensional bottleneck feature extracted from the Fisher trained bottleneck DNN.

We compare the results of MFCC feature and tandem feature in Table 3. Here the DNN model in DNN i-vector back-end is the RSR adapted DNN in section 4.2. From all the results based on MFCC feature, results of female task are obviously worse than results of male task. This maybe due to the relatively shorter vocal tract in females and the resulting higher formant frequencies in speech, which results in some speaker characteristics not reflected in MFCC feature. This observation is consistent with the phenomenon that people tend to be harder to distinguish among females' voice than among men's voice. However, when we use the tandem features in GMM i-vector, DNN i-vector and bottleneck DNN i-vector back-end, the gap between the performance of male and female task is consistently narrowed. Bottleneck features seem to make up for the shortage of MFCC features in the high frequency region. Besides, tandem features could consistently improve the performance of both male and female task compared to MFCC features. Furthermore, the improvement is more for GMM i-vector back-end (from 2.85% / 4.55% to 2.02% / 2.65%) than for DNN i-vector back-end (from 1.75% / 2.45% to 1.71% / 2.18%) since the phonetic awareness has already been very much exploited in DNN, frame-level, senone posteriors.

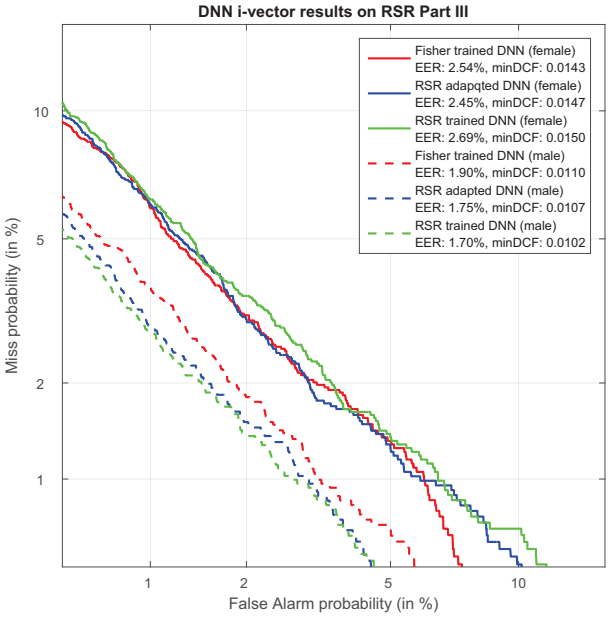

Figure 1: *Result comparison of DNN i-vector approach with different DNN models on RSR2015 Part III evaluation.*

Table 3: *Comparison of MFCC feature and tandem feature on RSR2015 Part III evaluation. The notation in EER means male / female.*

Model	Feature	EER (%)
GMM i-vector	MFCC feature	2.85 / 4.55
	Tandem feature	**2.02 / 2.65**
DNN i-vector (RSR adapted)	MFCC feature	1.75 / 2.45
	Tandem feature	**1.71 / 2.18**
Bottleneck DNN i-vector (Fisher trained)	MFCC feature	1.98 / 2.54
	Tandem feature	**1.81 / 1.84**

5. Conclusions

In recent years, applications of speaker recognition in telephone banking, smart home, artificial intelligence, etc. have led to more research attentions on text-dependent and text-constrained speaker verification. In this work, we investigate the use of DNN i-vector on text-constrained speaker verification using randomly prompted digit strings with RSR2015 Part III evaluation. We improve the EER performance by a relative decrease of 7.89% (from 1.90% to 1.75%) for male task and 3.54% (from 2.54% to 2.45%) for female task through DNN adaptation and digit senone selection. Besides, digit senone selection also compressed the model size of i-vector extractor for more than ten times. On the other hand, tandem features could not only improve the performance by using bottleneck feature to provide a complementary information for the acoustic MFCC feature, but also narrow the performance gap between male and female task by using bottleneck features to make up the shortage of MFCC features.

6. References

[1] D. A. Reynolds and W. M. Campbell, "Text-independent speaker recognition," in *Springer Handbook of Speech Processing*. Springer, 2008, pp. 763–782.

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

[3] 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.

[4] 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.

[5] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 711–720, 1997.

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

[7] Y. Lei, L. Ferrer, M. McLaren *et al.*, "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*. IEEE, 2014, pp. 1695–1699.

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

[9] T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann, and P. Dumouchel, "Text-dependent speaker recognition using PLDA with uncertainty propagation," *matrix*, vol. 500, p. 1, 2013.

[10] ——, "I-vector/PLDA variants for text-dependent speaker recognition," *preparation*, 2013.

[11] T. Stafylakis, M. J. Alam, and P. Kenny, "Text-dependent speaker recognition with random digit strings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1194–1203, 2016.

[12] 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.

[13] L. Chen, K.-A. Lee, B. Ma, W. Guo, H. Li, and L.-R. Dai, "Phone-centric local variability vector for text-constrained speaker verification." in *INTERSPEECH*, 2015, pp. 229–233.

[14] L. Chen, K. A. Lee, E.-S. Chng, B. Ma, H. Li, and L. R. Dai, "Content-aware local variability vector for speaker verification with short utterance," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5485–5489.

[15] W. Li, T. Fu, H. You, J. Zhu, and N. Chen, "Feature sparsity analysis for i-vector based speaker verification," *Speech Communication*, vol. 80, pp. 60–70, 2016.

[16] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, and S. Shamma, "Linear versus mel frequency cepstral coefficients for speaker recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 559–564.

[17] F. Grézl, M. Karafiát, S. Kontár, and J. Cernocky, "Probabilistic and bottle-neck features for LVCSR of meetings," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4. IEEE, 2007, pp. IV–757.

[18] A. K. Sarkar, C.-T. Do, V.-B. Le, and C. Barras, "Combination of cepstral and phonetically discriminative features for speaker verification," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1040–1044, 2014.

[19] B. H. Story, "Using imaging and modeling techniques to understand the relation between vocal tract shape to acoustic characteristics," in *Proc. Stockholm Music Acoustics Conf*. Citeseer, 2003, pp. 435–438.