COMBINING DEEP EMBEDDINGS OF ACOUSTIC AND ARTICULATORY FEATURES FOR SPEAKER IDENTIFICATION

Qian-Bei Hong[1], *Chung-Hsien Wu*[1,2], *Hsin-Min Wang*[1], *and Chien-Lin Huang*[3]

[1]Graduate Program of Multimedia Systems and Intelligent Computing,
National Cheng Kung University and Academia Sinica, Tainan, Taiwan
[2]Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan
[3]PingAn AI Lab, Palo Alto, CA 94306, USA

ABSTRACT

In this study, deep embedding of acoustic and articulatory features are combined for speaker identification. First, a convolutional neural network (CNN)-based universal background model (UBM) is constructed to generate acoustic feature (AC) embedding. In addition, as the articulatory features (AFs) represent some important phonological properties during speech production, a multilayer perceptron (MLP)-based AF embedding extraction model is also constructed for AF embedding extraction. The extracted AC and AF embeddings are concatenated as a combined feature vector for speaker identification using a fully-connected neural network. This proposed system was evaluated by three corpora consisting of King-ASR, LibriSpeech and SITW, and the experiments were conducted according to the properties of the datasets. We adopted all three corpora to evaluate the effect of AF embedding, and the results showed that combining AF embedding into the input feature vector improved the performance of speaker identification. The LibriSpeech corpus was used to evaluate the effect of the number of enrolled speakers. The proposed system achieved an EER of 7.80% outperforming the method based on x-vector with PLDA (8.25%). And we further evaluated the effect of signal mismatch using the SITW corpus. The proposed system achieved an EER of 25.19%, which outperformed the other baseline methods.

Index Terms— Speaker identification, convolutional neural network, articulatory feature

1. INTRODUCTION

Articulatory feature (AF) is an important representation of phonological properties during speech production. More precisely, AFs are abstract classes describing the movements or positions of different articulators during speech production. In the past years, AFs have been successfully used as features in speech recognition [1-4], but rarely investigated in speaker recognition.

Traditional text-independent speaker recognition systems typically used Gaussian mixture model-universal background model (GMM-UBM) and i-vector to represent the short-term spectral characteristics of speakers [5-6]. In recent years, deep neural network (DNN)-based models for speaker recognition have become more and more popular [7-9]. The most common method used the d-vector embedding to evaluate the performance of speaker recognition [10]. For improving the performance of speaker recognition, Waibel et al. [11] constructed a time delay neural network (TDNN) model that considered the temporal structure of acoustic events. Snyder et al. [12] indicated that TDNN outperformed traditional DNN-based methods. The x-vector systems based on TDNN structure has verified that the embedding representation was superior for short speech segments [13-14].

Most of the DNN-based methods used speaker embedding to compare with the embedding features of the speakers for recognition using probabilistic linear discriminant analysis (PLDA) backend [15]. McCree et al. [16] extended the PLDA model to include segment duration as well as to distinguish between session and channel variability. Rohdin et al. [17] developed an end-to-end speaker verification system that is initialized to mimic an i-vector with PLDA baseline. Snyder et al. [18] used a diarization system based on PLDA as a front-end for speaker recognition. On the other hand, Vincent et al. [19] indicated that the acoustic mismatches between training and testing data would cause performance degradation. The acoustic mismatch also occurred in the UBM training, speaker enrollment and utterance testing phases.

In summary, this paper integrated acoustic feature (AC) embedding with AF embedding for speaker identification. Adding the AFs is helpful for presenting the personal pronunciation attributes to improve speaker identification performance. Besides, using CNN-based feature embedding extraction followed by fully connected network, the proposed speaker identification can achieve a better performance compared to the traditional feature extraction methods.

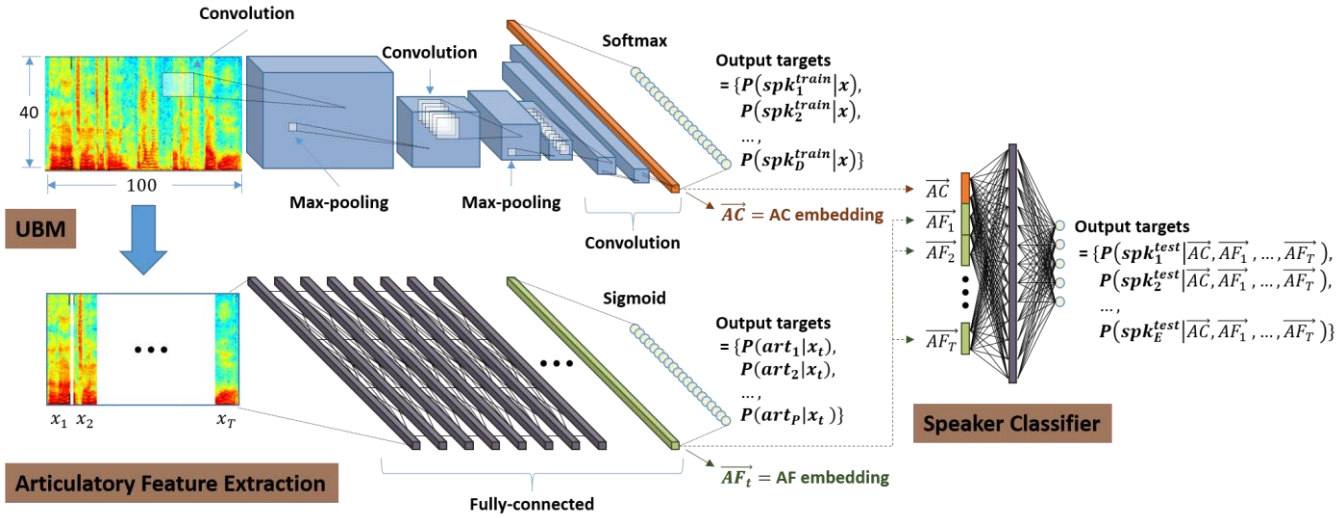


Fig. 1. The AC embedding and AF embedding classifier architecture for speaker identification

2. THE PROPOSED SPEAKER IDENTIFICATION SYSTEM

The proposed speaker identification system based on AC embedding and AF embedding is illustrated in Fig. 1. In this framework, a universal background model (UBM) is constructed for extracting the AC embedding. At the same time, the AF extractor is used to extract the AF embedding to capture the personal pronunciation features. The speaker classifier is trained by enrolled speakers' speech signals based on the latent speaker information in AC embedding and AF embedding.

2.1. Universal background model

In this study, a CNN-based UBM is trained to produce AC embedding. The UBM architecture consists of five convolution layers, two maximum pooling layers and a dense with softmax output layer. Each convolutional layer with batch normalization and Parametric Rectified Linear Unit (PReLU) [20] is used to increase the efficiency of model convergence in the training step. The whole layers of UBM is outlined in Table 1.

A 100×40 spectrogram is extracted from a 25ms window with a stride of 10ms to obtain a sequence of one-second long spectral features. The input features are the 40-dimentional Mel-frequency energy coefficients (MFECs) [21] for each sliding window. In the training stage, each spectrogram is fed to the UBM and turned into a probability vector. As the UBM converges, the well-trained UBM is used for further speaker feature extraction.

In training parameter setting, the weights of the convolution layers and the dense layer is initialized by normal distribution with zero-mean and a standard deviation of 0.1. The L2 regularization is used for weight decay, and the cost function utilizes the cross entropy for error estimation.

Table 1. The CNN-based UBM architecture

Layer	Layer Type	Kernel Size	Depth	Stride	Data Size
-	input	-	-	-	[100,40,1]
1	convolution	[1,5] [9,1]	16 32	[1,1] [2,1]	[46,36,32]
2	max-pooling	[2,2]	-	[2,2]	[23,18,32]
3	convolution	[1,5] [8,1]	32 64	[1,1] [1,1]	[16,14,64]
4	max-pooling	[2,2]	-	[2,2]	[8,7,64]
5	convolution	[1,3] [6,1]	128 128	[1,1] [1,1]	[3,5,128]
6	convolution	[1,3] [3,1]	256 512	[1,1] [1,1]	[1,3,512]
7	convolution	[1,3]	1024	[1,1]	[1,1,1024]
8	dense & softmax	-	-	-	[500]

2.2. Articulatory feature extraction

AFs can be distinguished into different pronunciation places and manners by speaker voices [22]. As shown in Table 2, 20 attributes, defined in Mandarin Chinese, are used to train an AF extractor.

As each speech phone corresponds to several attributes, in this study, the Kaldi ASR toolkit [23] is used to align the phone positions of the training speech signals in GMM-HMM-based acoustic model training procedure. According to the alignment information, every segment of training speech signals can exactly be labeled with the attributes which the phone corresponds to, and then used to train an AF extractor.

A multilayer perceptron (MLP)-based model is utilized for AF recognition, and the input features are constructed by 39-dimentional Mel-frequency cepstral coefficients (MFCCs) including delta and delta-delta features. In this model, the input spectrogram is extracted from a 0.1 second window with a stride of 0.05 second, which can present one kind of phone features in 0.1 second. This model consists of twenty-five hidden layers and each hidden layer was a 1,024-

dimentional feature with PReLU activation function. The mean squared error cost function is used to determine the loss of multi-target prediction.

Table 2. The mandarin speech attributes

Category	Attribute
Manner	Approximant, Fricative, Nasal, Stop, Vocalic
Place	Anterior, Back, Continuant, Coronal, Dental, High, Labial, Low, Mid, Retroflex, Round, Tense, Velar, Voiced
Silence	Silence

2.3. Embedding extraction

After UBM and AF extraction model training are completed, their embedding can be extracted from the models. Based on the assumptions of deep learning-based classifier, the last hidden layer output can be used as the classifiable features. In this study, the last convolutional output of UBM is regarded as an AC embedding and the last hidden layer output of AF extraction model is used as an AF embedding. As shown in Fig. 1, a one-second spectrogram can be divided into ten equal length sub-spectrograms; each sub-spectrogram is equivalent to 0.1 seconds long. The one AC embedding and ten AF embedding can be obtained from a one-second spectrogram. The AC and AF embedding are concatenated to form a combined embedding for further speaker classifier training.

2.4. Enrolled speaker classifier

The speaker classifier is trained to identify who is the speaker of the test recordings. All combined embeddings are obtained from the enrolled speakers' recordings, and the MLP-based classifier is used to extract the speaker information from the combined embedding. In this study, the 11,264-dimentional combined embedding is fed to this model, using one hidden layer to produce the 1,024-dimensional feature for dimensionality reduction. After that, the dense layer with softmax function produces the result of enrolled speakers' probabilities for speaker identification determination.

3. EXPERIMENTS AND RESULTS

3.1. Datasets

This study used three datasets to evaluate the proposed method, and the experiments were based on the analysis for three situations according to the properties of datasets.

- **King-ASR corpus:** The King-ASR-044 corpus is a Taiwanese speech recognition database collected from mobile devices [24]. This database contains the voices of 5,232 different speakers (2,365 males and 2,867 females), and each speaker was recorded in a quiet or noisy environment. This study used 500 randomly selected speakers for UBM training; the training data contained 15,396 recordings. On the other hand, the King-ASR-360 corpus is a Taiwanese and English Mixed Speech Recognition Corpus [24]. This corpus contains the voices of 1,026 different speakers (463 males, 563 females). This study used 2,082 randomly selected speakers from King-ASR-044 and 1,026 randomly selected speakers from King-ASR-360 for AF extraction model training. Finally, the 100 randomly selected speakers from King-ASR-044 were used for testing.
- **LibriSpeech corpus:** This is a corpus of approximately 1,000 hours of 16kHz read English speech. This study used the train-clean subset (totally 460 hours "clean" speech collected from 1,172 speakers) to test and analyze the speaker identification performance on different number of enrolled speakers.
- **Speakers in the Wild (SITW) corpus:** This corpus provides samples of nearly 300 individuals across clean interview, red carpet interviews, stadium conditions, outdoor conditions, and multi-speaker scenarios [25]. This study used the core-core subset (without interference from other speaker in the recordings, and a total of 1,201 recordings were collected from 180 speakers) to evaluate the speaker identification performance based on the same speaker across varying conditions.

3.2. Experimental setup

The dataset was labeled with speaker identity and the silence interval was removed using the energy-based voice activity detection. After that, all speech signals of the same speaker were concatenated into one speech signal and applied short-time Fourier transform method to extract the magnitude in frequency domain. This paper evaluated the speaker identification performance of all systems based on one-second long spectral features.

In this study, the methods using d-vector and x-vector were considered as the baseline systems. The d-vector embedding was obtained from UBM described in Section 2. The x-vector embedding was obtained from the same model in [14], the statistics pooling aggregates 100 frame-level outputs to form a segment-level vector of one second, and the embeddings were extracted at layer segment7 (last hidden layer). In addition, this study also compared the proposed method with cosine distance and PLDA backend in the aspect of scoring. All neural network models were implemented by TensorFlow.

3.3. The effect of AF embedding

Table 3 shows the results in terms of equal error rate (EER). As the experiments were based on speaker identification task, the prediction scores would be clustered by one-hot label (1: speaker, 0: impostor), to evaluate the speaker recognition performance in EER. Comparison on the LibriSpeech and SITW corpora, the classifier by concatenating AC embedding and AF embedding as input outperformed the embedding classifier only using the AC embedding. The

performance of the proposed method was worse compared to the methods using d-vector or x-vector for evaluation on King-ASR database. Contrast to the evaluation on King-ASR database, the proposed method performed better for the other two databases.

Table 3. EER (%) comparison on the baseline systems and the proposed system

Systems	King-ASR 100 speakers	LibriSpeech 1,172 speakers	SITW core-core 180 speakers
d-vector (cosine distance)	4.10	13.50	34.42
x-vector (cosine distance)	2.92	11.18	31.14
d-vector (PLDA)	1.54	10.61	31.43
x-vector (PLDA)	**1.02**	8.25	28.70
AC embedding classifier	2.33	7.95	26.67
AC & AF embedding classifier	2.41	**7.80**	**25.19**

3.4. Comparison on different number of enrolled speakers

In the experiment on King-ASR database, 100 randomly selected speakers were used for enrollment and evaluation. Five recordings from each speaker were selected randomly from the dataset to evaluate speaker identification performance, and the remaining containing 25 recordings per speaker were used to enroll in the baseline systems and train the speaker classifier. The proposed system achieved an EER of 2.41%; the baseline system using x-vector and PLDA achieved the best performance (1.02%).

The LibriSpeech corpus was used for enrollment and evaluation. 10 recordings from each of the 1,172 speakers was selected randomly from the database to enroll in the baseline systems and train the speaker classifier, and two recordings selected randomly from each of the 1,172 speakers were used to evaluate speaker identification performance. The proposed system achieved an EER of 7.80% outperforming the baseline systems. Fig. 2 illustrates the DET curves for comparison on 1,172 enrolled speakers.

As stated above, the proposed system outperformed the baseline systems when the number of enrolled speakers was increased, which was caused by the large number of data used to train the neural network model.

3.5. The effect of signal mismatch

The SITW corpus provides samples of same speaker across varying environmental conditions. The samples of the same speaker had different signal conditions even though those were recorded in the same environment. In this experiment, the signal mismatch problem was considered for enrollment and evaluation with different samples. Fig. 3 shows the DET curves for comparison on SITW core-core subset. One sample from each speaker was selected randomly from the dataset for evaluation, and the remaining samples were used to enroll in the baseline systems and train the speaker classifier. The results showed that all systems achieved poor EER, because the problem of signal mismatch significantly degraded the performance, and the proposed system achieved an EER of 25.19% but outperformed the baseline systems.

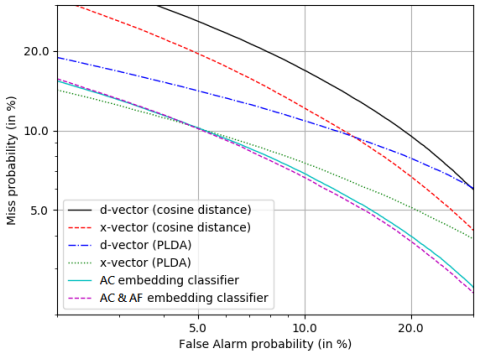


Fig. 2. DET curves comparison on 1,172 enrolled speakers

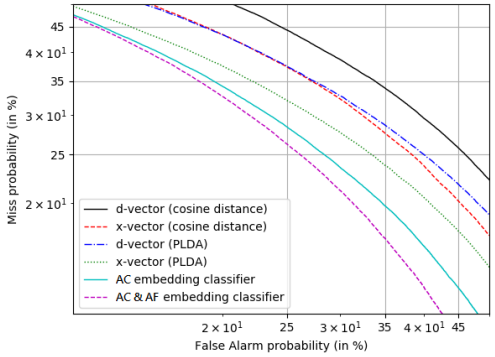


Fig. 3. DET curves comparison on SITW core-core subset without considering the types of environment

4. CONCLUSIONS

In this paper, we integrated AC embedding with AF embedding for speaker identification. An AF embedding can be used to present the personal pronunciation attributes to improve speaker identification performance. The experiments used the King-ASR series dataset to train the UBM and AF extraction model, and used the LibriSpeech corpus and SITW corpus to evaluate model performance. In the analysis of AF embedding, the experimental results on LibriSpeech and SITW corpora showed that the system concatenating AC embedding and AF embedding as input outperformed the embedding classifier only using AC embedding. Comparison on different number of enrolled speakers, the proposed system achieved the EER of 7.80% when the number of enrolled speakers was increased to 1,172 speakers, which outperformed the system using x-vector with PLDA. Finally, this study showed the analysis on signal mismatch problem by evaluating the performance on SITW corpus. The proposed system achieved the EER of 25.19% outperforming the other baseline systems.

5. REFERENCES

[1] V. Mitra, G. Sivaraman, C. BarteIs, H. Nam, W. Wang, C. Espy-Wilson, D. Vergyri, and H. Franeo, "Joint modeling of articulatory and acoustic spaces for continuous speech recognition tasks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 5205-5209, 2017.

[2] C. H. Wu, H. P. Shen and Y. T. Yang, "Chinese-english phone set construction for code-switching ASR using acoustic and DNN-extracted articulatory features," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 22, no. 4, pp. 858-862, 2014.

[3] G. Srinivasan, A. Illa, and P. K. Ghosh, "A study on robustness of articulatory features for automatic speech recognition of neutral and whispered speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 5936-5940, 2019.

[4] C. H. Wu, H. P. Shen, and C. S. Hsu, "Code-switching event detection by using a latent language space model and the delta-bayesian information criterion," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 23, no. 11, pp. 1892-1903, 2015.

[5] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "I-vector based speaker recognition on short utterances," in *INTERSPEECH*, pp. 2341-2344, 2011.

[6] A. Larcher, P.-M. Bousquet, K. A. Lee, D. Matrouf, H. Li, and J.-F. Bonastre, "I-vectors in the context of phonetically-constrained short utterances for speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 4773-4776, 2012.

[7] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances," in *INTERSPEECH*, pp. 1487-1491, 2017.

[8] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *INTERSPEECH*, pp. 999-1003, 2017.

[9] G. Bhattacharya, J. Alam, and P. Kenny, "Deep speaker embeddings for short-duration speaker verification," in *INTERSPEECH*, pp. 1517-1521, 2017.

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

[11] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K.J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 37, no. 3, pp. 328-339, 1989.

[12] D. Snyder, D. Garcia-Romero, and D. Povey, "Time delay deep neural network-based universal background models for speaker recognition," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 92-97, 2015.

[13] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *INTERSPEECH*, pp. 999-1003, 2017.

[14] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: robust DNN embeddings for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 5329-5333, 2018.

[15] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matějka, and N. Brümmer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 4832-4835, 2011.

[16] A. McCree, G. Sell, and D. Garcia-Romero, "Extended variability modeling and unsupervised adaptation for PLDA speaker recognition," in *INTERSPEECH*, pp. 1552-1556, August 2017.

[17] J. Rohdin, A. Silnova, M. Diez, O. Plchot, P. Matějka, and L. Burget, "End-to-end DNN based speaker recognition inspired by i-vector and PLDA," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 4874-4878, 2018.

[18] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 5796-5800, 2019.

[19] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech and Language*, vol. 46, pp. 535-557, 2017.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: surpassing human-level performance on imagenet classification," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 1026-1034, 2015.

[21] A. Torfi, N. M. Nasrabadi, and J. Dawson, "Text-independent speaker verification using 3d convolutional neural networks," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, 2018.

[22] C. H. Lee, and S. M. Siniscalchi, "An information-extraction approach to speech processing: Analysis, detection, verification, and recognition," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1089-1115, 2013.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, and J. Silovsky, *The Kaldi speech recognition toolkit*, No. CONF, IEEE Signal Processing Society, 2011.

[24] KingLine Data Center, *ASR-Corpus Commercial Resources*, [online] Available: https://kingline.speechocean.com/category.php?id=120, 2019

[25] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (SITW) speaker recognition database," in *INTERSPEECH*, pp. 818-822, 2016.