Sequential Speaker Embedding and Transfer Learning for Text-Independent Speaker Identification

Qian-Bei Hong[*], Chung-Hsien Wu[*†], Ming-Hsiang Su[†], and Hsin-Min Wang[*]
[*]Graduate Program of Multimedia Systems and Intelligent Computing,
National Cheng Kung University and Academia Sinica, Taiwan
[†]Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan
E-mail: {qbhong75, chunghsienwu, huntfox.su}@gmail.com, whm@iis.sinica.edu.tw

Abstract— **In this study, an approach to speaker identification is proposed based on a convolutional neural network (CNN)-based model considering sequential speaker embedding and transfer learning. First, a CNN-based universal background model (UBM) is constructed and a transfer learning mechanism is applied to obtain speaker embedding using a small amount of enrollment data. Second, considering the temporal variation of acoustic features in an utterance of a speaker, this study generates sequential speaker embedding to capture temporal characteristics of speech features of a speaker. Experiments were conducted on the King-ASR series database for UBM training, and the LibriSpeech corpus was adopted for evaluation. The experimental results showed that the proposed method using sequential speaker embedding and transfer learning achieved an equal error rate (EER) of 6.89% outperforming the method based on *x*-vector and PLDA method (8.25%). Furthermore, we considered the effect of speaker number for speaker identification. When the number of enrolled speakers was from 50 to 1172, the identification accuracy of the proposed method was degraded from 82.99% to 73.26%, which outperformed the identification accuracy of the method using *x*-vector and PLDA which was dramatically degraded from 83.17% to 60.95%.**

I. Introduction

Transfer learning has been used extensively to improve the reliability of features extraction [1] and provides the task extensibility [2], especially for acoustic model construction in automatic speech recognition (ASR) [3-4]. Speaker identification (SI) task is to identify the speaker from an utterance of a speaker by comparing the voice biometrics of the utterance with those speaker voice models stored beforehand. Speaker identification category can be divided into two subcategories: text-dependent and text-independent. In text-dependent SI, the speaker must utter the same phrases or words that are previously used for training while in text-independent SI, there is no constraint on the phrase or words. In this paper, we focused on text-independent speaker identification task.

In the past ten years, most SI methods were based on Gaussian mixture model-universal background model (GMM-UBM) [5-9]. Recently, deep neural network (DNN) architecture for speaker recognition has become more and more popular [10-13]. The DNN-UBM can improve the representation ability of speaker features using the deep structure. As embedding has been widely applied for presenting the recognizable features on deep learning [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]. As the embedding was obtained from the UBM trained from many speakers, the transfer learning mechanism [16] was applied to extract the embedding of a new speaker to increase the precision for speaker modeling.

On the other hand, Vincent et al. [17] and Huang et al. [18] indicated that the acoustic mismatches between training and testing data would cause performance degradation. The mismatch also occurred in the development and enrollment phases. The feature projection mechanism may not ensure that the characteristics of the enrolled speaker could be ideally projected to the space of the trained speakers. The time delay neural network (TDNN) [19] considered the temporal structure of acoustic events and outperformed traditional DNN-based methods [20]. The *x*-vector-based systems based on TDNN structure has verified that the embedding representation was superior for short speech segments [21-22]. In this paper, we use sequential embedding to capture the temporal characteristics in an utterance of a speaker.

In summary, this paper integrates the transfer learning mechanism which uses a small amount of enrollment data and the sequential embedding by extracting the temporal variation of acoustic features in an utterance of a speaker for text-independent speaker identification. With the transfer learning and sequential embedding features, the proposed method achieves a better performance compared to the traditional methods for speaker identification.

II. Related Work

Speaker identification consists of three phases: development, enrollment and evaluation. The development phase is the universal background model (UBM) training that uses a large amount of data to define the speaker manifold. The enrollment phase is that the new speaker is enrolled by specific information of the speaker to construct a speaker-dependent model. The evaluation phase uses the enrolled speaker models to decide who the speaker of the test utterance is. Motivated by the powerful feature extraction capability of

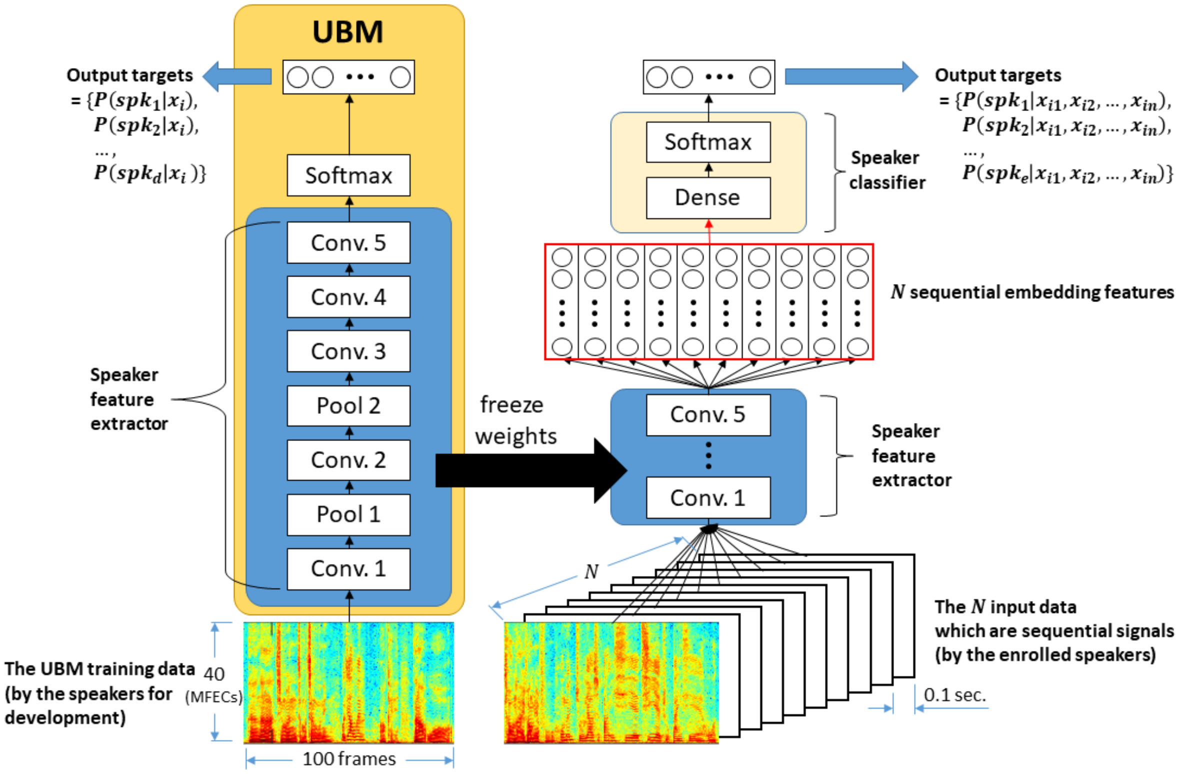
UBM
Output targets
$= \{P(spk_1|x_i),$
$P(spk_2|x_i),$
$\ldots,$
$P(spk_d|x_i)\}$
Softmax
Conv. 5
Conv. 4
Conv. 3
Pool 2
Conv. 2
Pool 1
Conv. 1
Speaker feature extractor
freeze weights
Softmax
Dense
Speaker classifier
Output targets
$= \{P(spk_1|x_{i1}, x_{i2}, \ldots, x_{in}),$
$P(spk_2|x_{i1}, x_{i2}, \ldots, x_{in}),$
$\ldots,$
$P(spk_e|x_{i1}, x_{i2}, \ldots, x_{in})\}$
N sequential embedding features
Conv. 5
Conv. 1
Speaker feature extractor
N
The N input data which are sequential signals (by the enrolled speakers)
0.1 sec.
The UBM training data (by the speakers for development)
40 (MFECs)
100 frames

Fig. 1 The sequential embedding DNN architecture for speaker identification.

DNNs, the DNN-based background model is used to directly model the speaker space. The *d*-vector-based SI system, in which the *d*-vector is created by averaging the last hidden layer outputs of the UBM, has become one of the most popular approaches in the adoption of DNN to speaker verification [10].

In addition to feature extraction, transfer learning technique has been efficiently used in task learning [23], image generation [24], speech recognition [25] and speaker verification [26]. The main concept of transfer learning is to share experience information from source domain to target domain. In speaker identification research, Jia et al. [2] used the embedding which was obtained by speaker identification model to synthesize speech. Their experimental results showed that the system significantly lowered the requirements for multi-speaker text-to-speech training data by separating the training of the speaker encoder and the synthesizer. Transfer learning is critical to achieve these results [2]. Zhang et al. [26] utilized transfer learning to solve the domain mismatched problem for speaker identification. The experimental results showed that better initial performance, faster speed of convergence, and better final performance were achieved for the neural network-based speaker verification framework with transfer leaning [26].

III. Sequential Speaker Embedding for Speaker Identification

A. Overview

The proposed speaker identification system is a two-stage DNN architecture, as depicted in Fig. 1. The first stage is UBM training for speaker feature extraction. The method is inspired by [27] which proposes a background model for adaptive feature learning. In this study, we consider spectrogram information to construct a CNN architecture to extract the generic speaker features. As the specific enrolled speakers are different from the speakers utilized to train the universal background model, the speaker features used to train the UBM could be transferred to the enrolled speakers. Next, the second stage is to train a DNN model to capture the temporal variation of speaker features over time. Finally, a dense layer with softmax produces the final likelihood for each speaker.

B. Spectral Feature Extraction

This study uses the 40-dimensional Mel-frequency energy coefficients (MFECs) [27] for each sliding window and obtains the spectral features of 100 frames per second as the input features of the UBM. The spectral features are extracted from a 25ms window with a stride of 10ms to obtain a sequence of spectral features. In addition, this study applies the voice activity detection (VAD) technique to detect the silence interval for silence removal. Silence removal is helpful to reduce the speaker identification error resulting from non-speech segments.

C. Universal Background Model

In this study, the CNN-based UBM consists of five convolution layers, two maximum pooling layers and a dense with softmax output layer. The size for each layer is illustrated in Table I. In addition, batch normalization, nonlinearity unit, and Parametric Rectified Linear Unit (PReLU) are applied to each convolutional output. The batch normalization is able to increase the efficiency of model convergence in the training step. PReLU proposed by He et al. [28] is the rectified activations in the convolutional network and is evaluated as better than ReLU in large scale image classification tasks.

In this study, we first set the spectrogram of size 100×40 as the input which represents the one-second spectral features. It is expected that one-second signal is enough to catch the characteristics of a phone from the speaker's utterance. Next, each input spectrogram is turned into an output vector that belongs to all training speaker's probabilities through layer transmission, abstraction and weighted sum. The whole architecture is a speaker classifier for the trained speakers. As the well-trained UBM is used for further speaker feature extraction, we need a large amount of data to train this model to cover all phonetic features.

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

TABLE I
THE CNN-BASED UBM ARCHITECTURE

Layer	Input size	Kernel size	Stride	Output size
Conv. 1	[100×40]	[1×5]×16 [9×1]×32	[1×1] [2×1]	[100×36]×16 [46×36]×32
Pool 1	[46×36]×32	[2×2]	[2×2]	[23×18]×32
Conv. 2	[23×18]×32	[1×5]×32 [8×1]×64	[1×1] [1×1]	[23×14]×32 [16×14]×64
Pool 2	[16×14]×64	[2×2]	[2×2]	[8×7]×64
Conv. 3	[8×7]×64	[1×3]×128 [6×1]×128	[1×1] [1×1]	[8×5]×128 [3×5]×128
Conv. 4	[3×5]×128	[1×3]×256 [3×1]×512	[1×1] [1×1]	[3×3]×256 [1×3]×512
Conv. 5	[1×3]×512	[1×3]×1024	[1×1]	[1×1]×1024
Softmax	[1×1]×1024	-	-	500

D. Speaker Embedding

After UBM training is completed, we remove the final dense layer and utilize the last convolutional output as the speaker embedding which represents a speaker's feature, similar to the *d*-vector. However, if there are many speakers enrolled in the speaker recognition system, the *d*-vector, which is the average of the last hidden layer outputs of the UBM, would face the problem of low discriminability. Different from *d*-vector which uses similarity comparison to make final decision, in this study, a sequence of speaker embedding features of the utterance is fed to a DNN-based classifier adopting transfer learning for speaker identification.

E. Sequential Embedding Classifier

Since the UBM is trained using the utterances of the speakers different from the enrolled speakers, speaker embedding representation using *d*-vector obtained by averaging the last hidden layer outputs of the UBM is unable to completely characterize the input speaker. In this study, we consider the sequential embedding representation from sequential signals to consider the temporal variation in the acoustic characteristics of an utterance produced by a speaker. The DNN-based classifier is used to extract the temporal relationship in sequential speaker embedding. Here, the length of input speech utterance is one second. The embedding is extracted and stacked into a 2-dimensional matrix along time axis. The interval between every two adjacent segments is 0.1 second. Finally, the input is a 10×1024 feature matrix. Then we produce the 1024-dimensional feature through the dense layer for dimensionality reduction. Equation (1) represents the element of 1024-dimmensional embedding from the outputs of the dense layer.

$$h_k = activation(\sum_{s=1}^{L\times N} w_{k,s} \cdot h_s^{ubm} + b_k) \tag{1}$$

where h_k is the *k*-th element obtained from the dense layer outputs, h_s^{ubm} is the *s*-th element of the sequential embedding obtained from the UBM, *L* is the number of elements in an embedding and *N* is the number of embeddings.

After that, the dense layer with softmax function produces the result of speakers' probabilities for speaker identification determination.

$$p(spk_e \mid x_{i1}, x_{i2}, ..., x_{in}) = soft\max(output_e) \tag{2}$$

$$output_e = \sum_{k=1}^{K} w_{e,k} \cdot h_k + b_e \tag{3}$$

where the speaker probability $p(spk_e \mid x_{i1}, x_{i2}, ..., x_{in})$ is obtained by giving a sequential spectrogram at time i.

IV. EXPERIMENTAL RESULTS

A. Experimental Databases

This study used two datasets, King-ASR-044 and LibriSpeech corpus, to train and evaluate the proposed method. King-ASR-044 [29] is a Taiwanese speech recognition database collected from mobile devices. This database contains the voices of 5,232 different speakers (2,365 males and 2,867 females) who are evenly distributed in age (16-30, 31-45, and >45), gender and regional accents. Each speaker was recorded in a quiet or noisy environment. This study used 500 speakers of the King-ASR-044 for UBM training. LibriSpeech corpus is an English speech database that is derived from read audiobooks. In this study, the LibriSpeech corpus was used for evaluating the performance of the proposed model.

In the first step, 500 randomly selected speakers were used for training the UBM. The training data contained 15,396 recordings. Then in the second step, clean signals of 460 hours from 1172 speakers in the LibriSpeech corpus were used for enrollment and evaluation. 10 recordings from each of the 1172 speakers was selected randomly from the database to enroll and train the speaker classifier, and two recordings selected randomly from each of the 1172 speakers were used to evaluate speaker identification performance. As mentioned above, this study used the unscripted voice samples to perform in every step, especially the enrollment phase. This will be close to the use in real environment.

B. *Experimental Setup*

The database was labeled with speaker identity and the silence interval was removed using the energy-based VAD. After that, all speech signals of the same speaker were concatenated as a single signal. The short-time Fourier transform was applied to extract the magnitude in frequency domain. The total lengths of the development, enrollment and evaluation signals were 13.13 hours, 32.45 hours and 5.33 hours, respectively. The average lengths of the enrollment and evaluation signals per speaker were about 100 seconds and 17 seconds. The spectrogram of the speech signal for one second was extracted and a shift of 0.1 second was applied to obtain a sequence of speech spectrograms as the experimental data. The 10 embedding features was combined to form a sequential embedding data. The total number of enrollment data was 1,168,358 and the total number of evaluation data was 192,000.

The study was implemented by TensorFlow. We set the training epochs to 100 with a mini-batch size of 128 and used the stochastic gradient descent algorithm to update the weights of the network. The learning rate was 0.05 with a decay factor of 0.94 which declined the learning rate per five epochs. Finally, an embedding dimensionality of 1024 was used for speaker identification.

C. *Performance Evaluation*

In this study, we aimed to use the transfer learning mechanism to improve the performance of speaker identification with a small-sized enrollment dataset. The methods using *x*-vector [21] was considered as the baseline system. The *x*-vector embedding was obtained from the same model in [21] with the 100 frame-level outputs for computing its mean and standard deviation, and the embeddings were extracted at layer segment7 (last hidden layer). The next two systems were transfer learning methods with different models for comparison. The DNN-based embedding model utilized the UBM to obtain speaker features, then used those speaker features directly to train a speaker classifier. The DNN-based sequential embedding model considered the temporal variation of speaker features by employing the ability of feature abstraction of DNN model to capture the speaker information. We evaluated the accuracy of those models based on equal error rate (EER).

In Table II, the experimental results showed that the sequential embedding DNN outperformed the embedding DNN and the baseline systems. The performance of using DNN-based classifier was better than the *x*-vector and PLDA method. Moreover, extracting the feature variation in the duration of one second could further improve the precision of identification. The detection error tradeoff (DET) curves were also used to evaluate the performance of *x*-vector, embedding DNN and sequential embedding DNN systems, as shown in Fig. 2. A DET graph is a graphical plot of error rates for classification systems, plotting the miss probability vs. false alarm probability. The results showed that the proposed sequential embedding DNN system outperformed either *x*-vector (cosine distance), *x*-vector (PLDA) or embedding DNN systems.

TABLE II
ACCURACY (%) AND EER (%) COMPARISON ON 1172 ENROLLED SPEAKERS

Method	Accuracy (%)	EER (%)
x-vector (cosine distance)	52.73	11.18
x-vector (PLDA)	60.95	8.25
Embedding DNN	66.82	7.95
Sequential embedding DNN	**73.26**	**6.89**

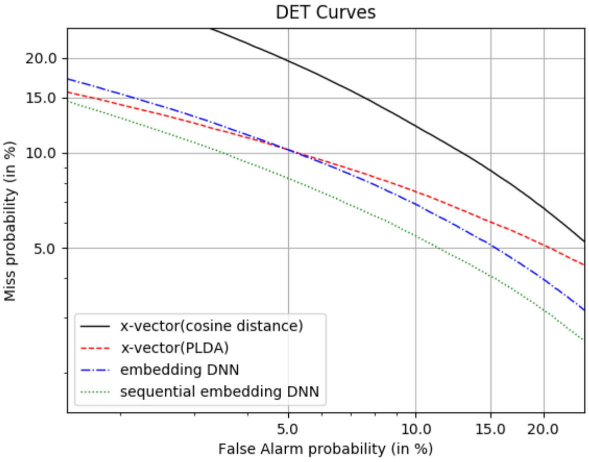

Fig. 2 DET curves comparison on 1172 enrolled speakers.

Table III shows the comparison results of the baseline systems and the proposed system on 50 enrolled speakers. Fig. 3 illustrated the DET curves for comparison on 50 enrolled speakers. It was a well-known fact that as the number of enrolled speakers was increased, the identification performance was likely to decrease. In this experiment, as the enrolled speakers increased from 50 to 1172, the identification accuracy of the sequential embedding DNN was only degraded from 82.99% to 73.26%, while the identification accuracy of the *x*-vector and PLDA system was dramatically degraded from 83.17% to 60.95%. Obviously, the proposed method outperformed the baseline systems when the number of speakers increased.

TABLE III
ACCURACY (%) AND EER (%) COMPARISON ON 50 ENROLLED SPEAKERS

Method	Accuracy (%)	EER (%)
x-vector (cosine distance)	80.54	11.41
x-vector (PLDA)	**83.17**	5.58
Embedding DNN	79.91	7.04
Sequential embedding DNN	82.99	**5.35**

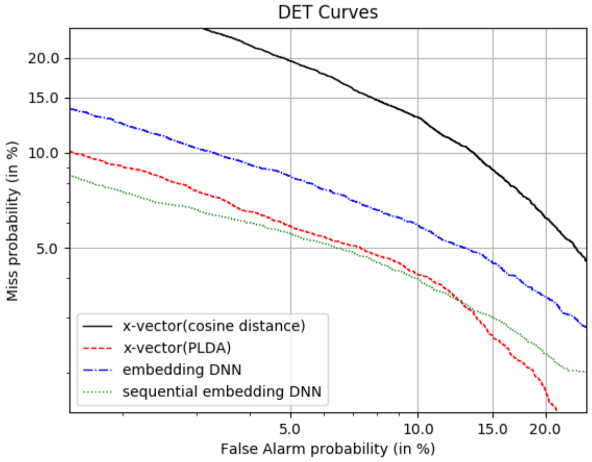

Fig. 3 DET curves comparison on 50 enrolled speakers.

V. Conclusions

In this paper, we integrated transfer learning with sequential embedding for speaker identification. The traditional similarity comparison method was replaced by the DNN classifier trained by the enrolled speakers' recordings. The experiments used the King-ASR series database to train the UBM and adopted the LibriSpeech corpus to evaluate model performance. The experimental results showed that the EER of sequential embedding DNN was 6.89% outperforming the method using *x*-vector and PLDA which achieved 8.25% EER. Then we considered the effect of different numbers of speakers and got the result that the sequential embedding DNN system achieved the best EER. In the further work, we will explore the influence of another mismatch issue such as environment mismatch.

References

[1] H. Lim, M. J. Kim, and H. Kim, "Cross-acoustic transfer learning for sound event classification," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Shanghai, China, Proceedings, pp. 2504-2508, March 2016.

[2] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. Lopez-Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in Neural Information Processing Systems*, Montréal Canada, Proceedings, pp. 4485-4495, December 2018.

[3] B. Pulugundla, M. K. Baskar, S. Kesiraju, E. Egorova, M. Karafiát, L. Burget, and J. Černocký, "BUT system for low resource Indian language ASR," in *INTERSPEECH*, Hyderabad, India, Proceedings, pp. 1-5, September 2018.

[4] J. Yi, J. Tao, Z. Wen, and Y. Bai, "Language-adversarial transfer learning for low-resource speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp.621-630, 2019.

[5] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308-311, 2006.

[6] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Toulouse, France, Proceedings, pp. I-97-I-100, May 2006.

[7] 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.

[8] N. Brummer, J. Cernocky, M. Karafiát, D. A. van Leeuwen, P. Matejka, P. Schwarz, A. Strasheim et al., "Fusion of heterogeneous speaker recognition systems in the stbu submission for the nist speaker recognition evaluation 2006," *IEEE Transactions on Audio Speech and Language Processing*, vol. 15, no. 7, pp. 2072-2084, 2007.

[9] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," *IEEE Transactions on Audio Speech and Language Processing*, vol. 15, no. 4, pp. 1448-1460, 2007.

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

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

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

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

[14] M. H. Su, C. H. Wu, K. Y. Huang, Q. B. Hong and H. M. Wang, "A chatbot using LSTM-based multi-layer embedding for elderly care," *IEEE International Conference on Orange Technology (ICOT)*, Singapore, Proceedings, pp. 70-74, December 2017.

[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)*, pp. 4832-4835, May 2011.

[16] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Transactions on Audio Speech and Language Processing*, vol. 21, no. 5, pp. 1060-1089, 2013.

[17] 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.

[18] K. Y. Huang, C. H. Wu, Q. B. Hong, M. H. Su, and Y. R. Zeng, "Speech emotion recognition using convolutional neural network with audio word-based embedding," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Taipei, Taiwan, Proceedings, pp. 265-269, November 2018.

[19] 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.

[20] 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)*, Scottsdale, AZ, USA, Proceedings, pp. 92-97, December 2015.

[21] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep neural network embeddings for text-independent speaker verification,” in *INTERSPEECH*, Stockholm, Sweden, Proceedings, pp. 999-1003, August 2017.

[22] 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)*, Calgary, AB, Canada, Proceedings, pp. 5329-5333, April 2018.

[23] C. Devin, A. Gupta, T. Darrell, P. Abbeel, and S. Levine, “Learning modular neural network policies for multi-task and multi-robot transfer,” in *IEEE International Conference on Robotics and Automation (ICRA)*, Singapore, Singapore, Proceedings, pp. 2169-2176, May 2017.

[24] D. Yoo, N. Kim, S. Park, A. Paek, and I. Kweon, “Pixel-level domain transfer,” in *European Conference on Computer Vision (ECCV)*, Amsterdam, The Netherlands, Proceedings, pp. 517-5322, October 2016.

[25] J. Kunze, L. Kirsch, I. Kurenkov, A. Krug, J. Johannsmeier, and S. Stober, “Transfer learning for speech recognition on a budget,” arXiv preprint, *arXiv preprint arXiv:1706.00290*, 2017.

[26] C. Zhang, S. Ranjan, and J. Hansen, “An analysis of transfer learning for domain mismatched text-independent speaker verification,” in *Odyssey 2018: The Speaker and Language Recognition Workshop*, Les Sables d’Olonne, France, Proceedings, pp. 181-186, June 2018.

[27] 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)*, San Diego, CA, USA, Proceedings, July 2018.

[28] 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)*, Santiago, Chile, Proceedings, pp. 1026-1034, December 2015.

[29] Taiwanese Speech Recognition Database (Mobile), 2019, [online] Available: http://kingline.speechocean.com/exchange.php?id=766&act=view