A Phone-Level Speaker Embedding Extraction Framework with Multi-Gate Mixture-of-Experts Based Multi-Task Learning

Zhijunyi Yang[1,3]*, Mengjie Du*[3,4]*, Rongfeng Su*[2,3]*, Xiaokang Liu*[3,4]*, Nan Yan*[2,3] *and Lan Wang*[2,3]

[1]School of Information Engineering, Wuhan University of Technology, Wuhan, China
[2]CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[4]University of Chinese Academy of Sciences, Beijing, China

yzjy@whut.edu.cn, {mj.du, rf.su, xk.liu, nan.yan, lan.wang}@siat.ac.cn

Abstract

Previous research has demonstrated that phonetic information can be incorporated into speaker embedding models to improve the performance of the speaker recognition system. However, the current speaker recognition process base on phonetic information is not in the line with the forensic voice identification process, where the identity of a speaker is determined by comparing confident phones one by one. This mainly manifests in two aspects. Firstly, real-world speech signals containing unreliable noisy data are directly used as model inputs. Secondly, the traditional utterance-level speaker embedding consisting of a single vector is a coarse-grained representation of speaker identity. To address these issues, this paper proposes a phone-level speaker embedding extraction framework with Multi-gate Mixture-of-Experts (MMoE) based multi-task learning. In the proposed framework, confident mono-phone segments are obtained by a well-trained ASR model and used as the inputs of the speaker embedding extractor with residual structure and self-attention mechanism. MMoE in the framework is supposed to increase the modeling sensitivity of different accents. To improve the fineness of the speaker representation, a vector set generated from different mono-phone segments is used as the phone-level speaker embedding. On a large-scale Mandarin speaker identification dataset, the proposed system using phone-level speaker embeddings and the MMoE technique significantly outperformed the utterance-level baseline system by the Top-1 average error rate reduction of 47.0% relatively.

Index Terms: speaker identification, phone-level speaker embedding, phonetic information, multi-task learning

1. Introduction

Speaker recognition can be classified into speaker identification and speaker verification. This paper will focus on speaker recognition techniques applied in speaker identification tasks that aim to determine an unknown speaker from a group of enrolled speakers. Recently, motivated by the powerful feature extraction capability of deep neural networks, lots of speaker recognition methods have been proposed to extract the speaker embeddings as the representation of speakers [1, 2, 3]. Such speaker embeddings can be applied in speaker identification or speaker verification tasks by using different scoring functions [4, 5, 6]. However, due to the diverse accents and pronunciations, large uncertainty exists in the acquired speaker embeddings and the system performance is degraded.

To reduce the uncertainty in speaker embeddings, various phonetic information based speaker recognition methods have been proposed. The most commonly used method is to extract the phonetic features derived from the bottleneck layer of an automatic speech recognition (ASR) model for the speaker embedding model training [7, 8]. Furthermore, to reduce noise effects in frame-level multi-task learning, a hybrid multi-task learning at frame and segment levels is employed in the speaker embedding model using phonetic information [9]. Inspired by traditional GMM-UBM methods that model each phoneme with an individual GMM, a phoneme-unit-specific time-delay neural network (TDNN) is proposed in [10]. In the phoneme-unit-specific TDNN, different TDNNs are trained for different phonemes and the final utterance-level speaker embeddings are obtained by the weighted sum of the phoneme posterior probabilities. Previous studies [7-10] have shown that additional phonetic information in the modeling stage can reduce the uncertainty in speaker embeddings and thus improve the speaker recognition system performance.

However, there are still two important issues associated with previous phonetic information based speaker recognition methods. Firstly, the real-world audio inputs contain lots of "noisy data", such as silence and unreliable voice segments. These "noisy data" would introduce additional variability during model training and thus increase the uncertainty in speaker embeddings. Secondly, traditional utterance-level speaker embedding consists of a single vector but contains diverse pronunciation characteristics for a specific speaker. The use of the coarse-grained utterance-level speaker embeddings for speaker identification tasks is not in the line with the forensic voice identification process [11], where the identity of a speaker is determined by comparing phones one by one.

To address these two issues, this paper proposes a phone-level speaker embedding extraction framework with Multi-gate Mixture-of-Experts (MMoE) based multi-task learning. In the proposed framework, confident mono-phone segments are obtained by a well-trained ASR model, while unreliable "noisy data" is discarded. Such confident mono-phone segments are used as the inputs of the speaker embedding extractor with residual structure and self-attention mechanism. Inspired by the success of the MMoE based multi-task learning in the field of recommendation systems [12], MMoE in the framework is supposed to increase the sensitivity of the model to different accents. In addition, to improve the fineness of the speaker representation, a vector set generated from different mono-phone segments is used as the final phone-level speaker embedding.

The main contributions of this paper are listed as follows:

- A phone-level speaker embedding extraction framework is proposed in this paper. Compared with traditional utterance-level speaker embeddings, the proposed phone-level speaker embeddings have a more refined characterization of the speaker identity.
- This paper investigates the combination of the MMoE

979-8-3503-9796-3/22/$31.00

technique with confident mono-phone segments and their application in the domain of speaker identification.
- This paper provides a phone-level speaker recognition system for large-scale Mandarin speaker identification tasks and it significantly outperformed the traditional utterance-level baseline system.

The rest of this paper is organized as follows. Section 2 describes the squeeze-and-excitation residual network (SE-ResNet) for extracting traditional utterance-level speaker embeddings. The phone-level speaker embedding extraction framework with MMoE is proposed in section 3. Experimental results will be presented in Section 4. The last section concludes and discusses possible future work.

2. Utterance-level Speaker Embedding Extraction

Squeeze-and-excitation residual network (SE-ResNet) is one of the state-of-the-art network architectures for speaker embedding extraction [13, 14, 15]. The Squeeze-and-Excitation (SE) module in SE-ResNet can adaptively recalibrate channel-wise feature responses by explicitly modeling interdependencies between channels [16, 17].

As shown in Table 1, the inputs of the SE-ResNet used in this paper are 64-dimensional static mel-filter bank (Fbank) with delta and double-delta coefficients. The SE-ResNet is composed of several stacked residual blocks with an SE module, which is referred as SE-ResBlocks. The "SE-Res64", "SE-Res128", "SE-Res256" and "SE-Res512" shown in Table 1 are the SE-ResBlocks with different settings.

After several SE-ResBlocks, the frame-level feature maps are integrated into an utterance-level vector through the self-attention pooling (SAP) mechanism [18]. To deal with the variable-length inputs, an additional mask is adopted in the "SAP" layer to remove the zero-padding values. Supposed that $X = [x_1, x_2, ..., x_N]$ is the output matrix of "SE-Res512" layer in Table 1, where x_i is the i-th time dimension matrix of X, the output of "SAP" layer is the weighted matrix U.

$$Q_i, K_i, V_i = (W^Q, W^K, W^V)x_i \tag{1}$$

$$Attention(Q_i, K_i, V_i) = softmax(mask(\frac{Q_i K_i^T}{\sqrt{s}}))V_i \tag{2}$$

$$U = \sum_i^N Attention(Q_i, K_i, V_i) \tag{3}$$

where Q_i, K_i, V_i are the linear transformation matrics of x_i, s is the scaling parameters (512 in this paper). $mask(\cdot)$ replaces the zero-padding value in x with negative infinity, so that the weight is zero after $softmax$ processing [19]. In the "Output" layer, softmax is used as the activation function, while cross-entropy is used as the loss function.

In the enrollment and test stages, the final "Output" layer in Table 1 will be removed and the utterance-level speaker embeddings are obtained from the outputs of the "Embedding" layer. The cosine function is used to measure the similarity between two different speaker embeddings.

3. Phone-level Speaker Embedding Extraction Framework

3.1. Phone-level input acquisition

In the forensic voice identification process, voiceprint experts usually extract the relatively clean speech segments according to different phones, and then confirm the identity of the criminal suspect by manually comparing the spectrogram of the acquired clean speech segments. According to this process, the real-world audio data containing unreliable "noisy data" need to be pre-processed before model training. As shown in Figure 1(b), an energy-based voice activity detector (VAD) is used to filter out long silence in the original speech signal. After that, confident mono-phone segments are obtained by using a well-trained ASR model. The output labels of the ASR model are tied tri-phones, while the mono-phone outputs are obtained by the intermediate phones of the tied tri-phones. For example, the corresponding mono-phone of the tri-phone "b-a3+w" is "a", where the tone is not considered in this paper. The toneless Mandarin phone set from Cambridge University [20] is used in this paper. This phone set consists of 44 mono-phones plus the silence ("sil") and the short pause ("sp").

Table 1: *Architecture of SE-ResNet. "SAP" denotes the self-attention pooling layer, "B" denotes the batch size, "T" denotes the number of short-time frames.*

Layer	Structure	Stride	Output size
Conv2d-64	$1 \times 1, 64$	2×1	$(B, 64, 32, T)$
SE-Res64	$\begin{pmatrix} Conv, 3 \times 3, 64 \\ Conv, 3 \times 3, 64 \\ fc, [8, 64] \end{pmatrix}$	1×1	$(B, 64, 32, T)$
Conv2d-128	$1 \times 1, 128$	2×2	$(B, 128, 16, T/2)$
SE-Res128	$\begin{pmatrix} Conv, 3 \times 3, 128 \\ Conv, 3 \times 3, 128 \\ fc, [16, 128] \end{pmatrix}$	1×1	$(B, 128, 16, T/2)$
Conv2d-256	$1 \times 1, 256$	2×1	$(B, 256, 8, T/2)$
SE-Res256	$\begin{pmatrix} Conv, 3 \times 3, 256 \\ Conv, 3 \times 3, 256 \\ fc, [32, 256] \end{pmatrix}$	1×1	$(B, 256, 8, T/2)$
Conv2d-512	$1 \times 1, 512$	2×1	$(B, 512, 4, T/2)$
SE-Res512	$\begin{pmatrix} Conv, 3 \times 3, 512 \\ Conv, 3 \times 3, 512 \\ fc, [64, 512] \end{pmatrix}$	1×1	$(B, 512, 4, T/2)$
SAP	-	-	$(B, 512)$
Embedding	$fc, [512, 512]$	-	$(B, 512)$
Output	$fc, [512, \#Spk]$	-	$(B, \#Spk)$

To acquire confident mono-phone segments, the confident metric should be defined. Supposed that m_i is the i-th mono-phone and x_t is the t-th frame, the confident score $c_{i,t}$ of x_t with respect to m_i is defined as:

$$c_{i,t} = p(m_i|x_t) \tag{4}$$

where $p(\cdot)$ is the posterior probability from a well-trained ASR model. The mono-phone output of the t-th frame is determined by using the maximum a posterior(MAP) criterion:

$$\tilde{m}(t) = \underset{i}{argmax}(c_{i,t}) = \underset{i}{argmax}(p(m_i|x_t)) \tag{5}$$

The consecutive frames with identical mono-phone outputs are concatenated into a mono-phone segment. The final confident mono-phone segments are selected through a pre-defined threshold $\widetilde{c}$. In addition, the silence segments ("sp" and "sil") are considered to be "noisy data" and discarded in this paper.

As shown in Figure 1(b), after the data pre-processing, the original audio inputs are cut into multiple mono-phone segments with corresponding mono-phone labels. As shown in Figure 1(a), the acquired confident mono-phone segments are used as the inputs of the speaker embedding extractor, while the mono-phone labels are used for subsequent multi-task learning.

3.2. MMoE based multi-task learning

The complexity and diversity of the speech content in the modeling process greatly increase the uncertainty of the speaker embedding. It will degrade the accuracy of the speaker recognition

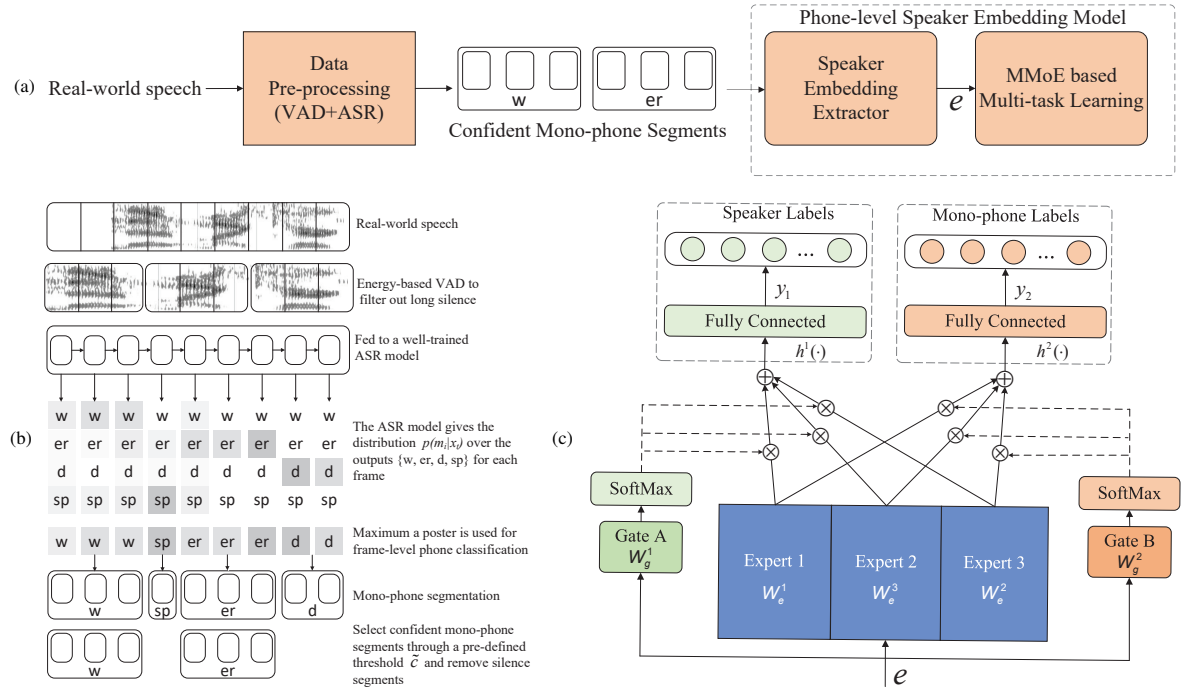

Figure 1: *(a) Overall of the proposed phone-level speaker embedding extraction framework. The "speaker embedding extractor" can be any neural network according to different requirements. In this paper, it was the SE-ResNet without the "Output" layer in Table 1. (b)Data pre-processing. (c) Multi-gate Mix-of-Experts (MMoE) based multi-task learning.*

system. In fact, when humans distinguish different speakers, they often get determination according to the inherent pronunciation characteristics of some speakers. Therefore, it is necessary to use the multi-task learning technique to enhance the phone-awareness of the model to find discriminative phones [21].

Traditional multi-task learning fully shares the bottom network parameters [22], which is equivalent to assuming that different sub-tasks have the same data distribution. However, the feature representations of different mono-phones and speakers should have different distributions. Therefore, the invalid assumption will greatly weaken the characterization of the final speaker embeddings. Inspired by the success of MMoE in the field of recommendation systems, MMoE based multi-task learning technique for speaker embedding extraction is proposed in this paper. Supposed that the output of the "Embedding" layer with respect to a mono-phone segment is the vector e, the MMoE shown in Figure 1(c) can be described as:

$$y_k = h^k(\sum_{i=1}^{3} gate^k(e)expert^i(e)), k = 1, 2 \quad (6)$$

$$gate^k(e) = softmax(W_g^k e), k = 1, 2 \quad (7)$$

$$expert^i(e) = W_e^i e, i = 1, 2, 3 \quad (8)$$

where h^k represents the sub-task networks. $gate^k$ denotes the gating networks, $expert^i$ denotes the expert networks.

As shown in Figure 1(c), each expert is regarded as a weak learner. According to the integration idea, the combination of several weak learners can become a strong learner. In addition, a gating network is used for each task. The gating networks take the input features and output softmax gates assembling the experts with different weights, allowing different tasks to utilize experts differently [12]. The outputs of the assembled experts are then passed into a speaker and mono-phone specific sub-networks. In a word, MMoE automatically adjusts the shared parameters and the unshared parameters, thus it can learn the relationship between different tasks.

3.3. Implement details

The overall of the proposed framework is shown in Figure 1(a), where the speaker embedding extractor can be replaced by a suitable network structure according to different requirements. In this paper, the SE-ResNet without the "Output" layers in Table 1 is used as the speaker embedding extractor shown in Figure 1(a). In the enrollment stage, the fine-grained phone-level speaker embedding for a specific speaker consists of a set of vectors obtained from confident mono-phone segment inputs [23]. Likewise, in the test stage, different mono-phone segments can generate different feature representations.

Supposed that the phone set of 44 mono-phones is Ω, the speaker embedding of the i-th speaker in the enrollment stage is $E^i = \{e^i_{m_k} | m_k \in M_k\}$, where $M_K = \{m_k | k = 1, 2, ..., K\} \subseteq \Omega$, K is the number of confident mono-phones of the i-th speaker, $e^i_{m_k}$ is the representation vector of the i-th speaker with respect to the mono-phone "m_k". Similarly, for a given test utterance with respect to an unknown speaker, supposed that the speaker embedding is $E' = \left\{e_{m'_j} | m'_j \in M'_{K'}\right\}$, where $M'_{K'} = \{m'_j | j = 1, 2, ..., K'\} \subseteq \Omega$, K' is the number of mono-phones in the test utterance, $e_{m'_j}$ is the representation vector of the unknown test speaker with respect to the mono-phone "m'_j", we can get the intersection $M_N = \{m''_n | n = 1, 2, .., N\} = M_K \cap M'_{K'}$ of M_K and $M'_{K'}$, where m''_n is the n-th confident mono-phone of both the i-th enrolled speaker and the unknown test speaker. Finally, the similarity of these two speakers $S(E^i, E')$ can be calculated as:

$$S(E^i, E') = \frac{1}{N} \sum_{n=1}^{N} cos(e^i_{m''_n}, e'_{m''_n}) \quad (9)$$

According to the selection of confident mono-phone segments, the value of N may be so small that the final similarity calculation may be less reliable. Therefore, this paper only considers the case of $N \geq 10$ (when confident threshold "$\tilde{c}$" $= 0.6$).

4. Experiments

4.1. Experimental setup

To investigate the performance of the phone-level speaker embedding, experiments were conducted on a large-scale Mandarin speaker identification dataset. The **training set** includes 20,000 speakers of 720 hours. The test set contains 2 subsets: (1) **"Seen" Set**: 16,672 speakers that occur in the **training set**. (2) **"Unseen" set**: 16,652 speakers that do not overlap with the **training set**. Note that there is no overlap between enrollment and test utterances. In both "Seen" and "Unseen" sets, each speaker has about 3 minutes speech data in the enrollment stage and about 30 seconds speech data in the test stage.

The ASR model obtained from 564 hours of broadcast news speech data [25] was used for extracting confident mono-phone segments. All experiments were implemented with Pytorch [26]. The models were trained using 4 NVIDIA A6000 GPU with 48GB memory for 20 epochs. The minibatch size was 128. We used the Adam optimizer with an initial learning rate of 0.001 decreasing by 5% every epoch [15]. The speaker identification systems shown in Table 2 are: ① the utterance-level baseline system without multi-task learning; ② the utterance-level system with traditional multi-task learning; ③ the system using confident mono-phone segments as inputs; ④ the system using confident mono-phone segments and traditional multi-task learning technique; ⑤ the proposed system shown in Figure 1(a).

4.2. Result analysis

Table 2: *The performance of various speaker identification systems measured by Top-1 error rate (Err), "$\widetilde{c}$" denotes the confident threshold in section 3.1. "Traditional" indicates the traditional multi-task learning [9] method and "MMoE" denotes the proposed MMoE structure in Figure 1(c).*

Systems	Embedding	Multi-task learning	Err(%)		
			Seen	Unseen	Avg
①	Utterance-level	-	3.33	3.42	3.38
②		Traditional	2.15	2.26	2.21
③	Phone-level ($\widetilde{c} = 0.6$)	-	2.51	2.50	2.50
④		Traditional	2.02	2.20	2.11
⑤		MMoE	**1.69**	**1.89**	**1.79**

The performance of various utterance-level and phone-level speaker identification systems with or without multi-task learning are shown in Table 2. From those results in Table 2, we found 4 major trends.

Phone-level vs utterance-level: Performance improvements can be obtained by using phone-level inputs. For example, comparing system ① with system ③ in Table 2, the phone-level speaker identification system outperformed the traditional utterance-level speaker identification system by a Top-1 average error rate reduction of 26% relatively. The results of system ② and ④ with multi-task learning also supported this point.

With or without multi-task learning: Comparing system ① with system ② (system ③ with system ④), we found that the multi-task learning using additional phonetic information as supervised labels can significantly improve the system performance in both phone-level and utterance-level cases.

Traditional multi-task learning vs MMoE based multi-task learning: As expected, the proposed system ⑤ using MMoE based multi-task learning outperformed system ④ using the traditional multi-task learning by a Top-1 average error rate reduction of 15.2% relatively.

Best system vs baseline system: Using all techniques including confident phone-level inputs described in section 3.1 and the MMoE technique proposed in section 3.2, the acquired phone-level system ⑤ gave the lowest Top-1 error rate and outperformed the traditional utterance-level baseline system ① by 47.0% relative on average.

Table 3: *The performance of the proposed phone-level system with MMoE based multi-task learning using different confident thresholds, where "0" means no threshold was set.*

Confident threshold ($\widetilde{c}$)	Err(%)		
	Seen	Unseen	Avg
0	2.67	2.90	2.78
0.5	1.97	2.05	2.01
0.6	**1.69**	**1.89**	**1.79**
0.7	1.81	2.04	1.92
0.8	2.04	2.45	2.25

In addition, the effect of using different confident thresholds on the proposed phone-level speaker identification system with MMoE based multi-task learning is shown in Table 3. Firstly, as shown in the first line of Table 3, when there is no confident threshold limitation, unreliable "noisy data" is retained in the speech inputs, which would introduce additional variability. Secondly, by setting suitable confident thresholds (from line 2 to line 5), we filter out "noisy data" and obtain reliable mono-phone segments as input, which can effectively improve system performance. Thirdly, when the threshold is too high, only a few mono-phone segments would be obtained from a real-world speech signal, which would lead to unrobust similarity scoring from Equation (9) and degrade the system performance.

5. Conclusions and future works

In this paper, inspired by the forensic voice identification process, we proposed a phone-level speaker embedding extraction framework using confident mono-phone segments and MMoE based multi-task learning. Benefiting from the phone-level speaker embeddings, the problems of unreliable "noisy data" and the coarse granularity of traditional utterance-level speaker embeddings were further addressed. The proposed framework was evaluated on a large-scale Mandarin speaker identification dataset. From the results shown in this paper, we concluded that confident phone-level inputs can significantly improve the system performance. MMoE based multi-task learning using phonetic information as supervised labels could further improve the performance of speaker identification systems. Moreover, a suitable confident threshold was also an important factor affecting the robustness of the speaker identification systems. The drawback of the proposed framework is the requirement of phoneme-rich speech data, which can ensure that enough confident mono-phone segments are available in both enrollment and test stages. In future work, we will verify the proposed framework on other benchmark datasets and combine it with traditional utterance-level speaker embedding extraction methods.

6. Acknowledgements

This work was supported in part by the National Key R&D Program of China (2020YFC2004100), National Natural Science Foundation of China (NSFC 61771461), Shenzhen Peacock Team Project (Grant No. KQTD20200820113106007) and Shenzhen Science and Technology Program (Grant No. JCYJ20210324115810030).

7. References

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

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

[3] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.

[4] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," in *Proceedings of The Speaker and Language Recognition Workshop (Odyssey)*, 2010, p. paper 14.

[5] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "Plda for speaker verification with utterances of arbitrary duration," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7649–7653.

[6] N. Dehak, R. Dehak, J. R. Glass, D. A. Reynolds, P. Kenny *et al.*, "Cosine similarity scoring without score normalization techniques." in *Proceedings of The Speaker and Language Recognition Workshop (Odyssey)*, 2010, pp. 15–19.

[7] T. Zhou, Y. Zhao, J. Li, Y. Gong, and J. Wu, "Cnn with phonetic attention for text-independent speaker verification," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 718–725.

[8] I. Viñals, D. Ribas, V. Mingote, J. Llombart, P. Gimeno, A. Miguel, A. O. Giménez, and E. Lleida, "Phonetically-aware embeddings, wide residual networks with time-delay neural networks and self attention models for the 2018 nist speaker recognition evaluation." in *Proceedings of ISCA INTERSPEECH*, 2019, pp. 4310–4314.

[9] Y. Liu, L. He, J. Liu, and M. T. Johnson, "Speaker embedding extraction with phonetic information," *Proceedings of ISCA INTERSPEECH*, pp. 2247–2251, 2018.

[10] X. Chen and C. Bao, "Phoneme-unit-specific time-delay neural network for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1243–1255, 2021.

[11] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[12] J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi, "Modeling task relationships in multi-task learning with multi-gate mixture-of-experts," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 1930–1939.

[13] H. Zhang, L. Wang, Y. Zhang, M. Liu, K. A. Lee, and J. Wei, "Adversarial separation network for speaker recognition." in *Proceedings of ISCA INTERSPEECH*, 2020, pp. 951–955.

[14] Z. Wang, F. Xu, K. Yao, Y. Cheng, T. Xiong, and H. Zhu, "Antvoice neural speaker embedding system for ffsvc 2020." in *Proceedings of ISCA INTERSPEECH*, 2021, pp. 1069–1073.

[15] J. S. Chung, J. Huh, S. Mun, M. Lee, H.-S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition." in *Proceedings of ISCA INTERSPEECH*, 2020, pp. 2977–2981.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[17] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[18] Y. Chen, L. Ren, H. Xia, Z. Wang, C. Gao, and F. Wang, "A compound fault diagnosis method based on multi-task learning with multi-gate mixture-of-experts," in *Proceedings of IEEE International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, 2022, pp. 281–285.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of Conference on Neural Information Processing Systems (NIPS)*, 2017, pp. 6000–6010.

[20] M. J. Gales, B. Jia, X. Liu, K. C. Sim, P. C. Woodland, and K. Yu, "Development of the cuhtk 2004 mandarin conversational telephone speech transcription system," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2005, pp. 841–844.

[21] N. Tawara, A. Ogawa, T. Iwata, M. Delcroix, and T. Ogawa, "Frame-level phoneme-invariant speaker embedding for text-independent speaker recognition on extremely short utterances," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6799–6803.

[22] Y. Liu, Z. Li, L. Li, and Q. Hong, "Phoneme-Aware and Channel-Wise Attentive Learning for Text Dependent Speaker Verification," in *Proceedings of ISCA INTERSPEECH*, 2021, pp. 101–105.

[23] D. Tan and T. Lee, "Fine-ed Style Modeling, Transfer and Prediction in Text-to-Speech Synthesis via Phone-Level Content-Style Disentanglement," in *Proc. Interspeech 2021*, 2021, pp. 4683–4687.

[24] S. Wang, J. Rohdin, L. Burget, O. Plchot, Y. Qian, K. Yu, and J. Cernockỳ, "On the usage of phonetic information for text-independent speaker embedding extraction." in *Proceedings of ISCA INTERSPEECH*, 2019, pp. 1148–1152.

[25] R. Su, X. Liu, L. Wang, and J. Yang, "Cross-domain deep visual feature generation for mandarin audio–visual speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 185–197, 2019.

[26] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, *PyTorch: An Imperative Style, High-Performance Deep Learning Library*, 2019, pp. 12–23.