MFA-Conformer: Multi-scale Feature Aggregation Conformer for Automatic Speaker Verification

Yang Zhang[1,2], Zhiqiang Lv[2], Haibin Wu[3,4], Shanshan Zhang[2], Pengfei Hu[2],
Zhiyong Wu[1,4,†], Hung-yi Lee[3], Helen Meng[4]

[1] Shenzhen International Graduate School, Tsinghua University
[2] TEG AI, Tencent Inc
[3] Graduate Institute of Communication Engineering, National Taiwan University
[4] Centre for Perceptual and Interactive Intelligence, The Chinese University of Hong Kong

`zhangy20@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn`

Abstract

In this paper, we present Multi-scale Feature Aggregation Conformer (MFA-Conformer), an easy-to-implement, simple but effective backbone for automatic speaker verification based on the Convolution-augmented Transformer (Conformer). The architecture of the MFA-Conformer is inspired by recent state-of-the-art models in speech recognition and speaker verification. Firstly, we introduce a convolution subsampling layer to decrease the computational cost of the model. Secondly, we adopt Conformer blocks which combine Transformers and convolution neural networks (CNNs) to capture global and local features effectively. Finally, the output feature maps from all Conformer blocks are concatenated to aggregate multi-scale representations before final pooling. We evaluate the MFA-Conformer on the widely used benchmarks. The best system obtains 0.64%, 1.29% and 1.63% EER on VoxCeleb1-O, SITW.Dev, and SITW.Eval set, respectively. MFA-Conformer significantly outperforms the popular ECAPA-TDNN systems in both recognition performance and inference speed. Last but not the least, the ablation studies clearly demonstrate that the combination of global and local feature learning can lead to robust and accurate speaker embedding extraction. We have also released the code[1] for future comparison.

Index Terms: speaker verification, Transformer, Conformer, speaker recognition

1. Introduction

Automatic speaker verification (ASV) is a task to verify whether a given utterance is from a claimed enrolled speaker. In recent years, it is witnessed the significant developments of ASV [1–6] and now ASV is a well-developed technology and widely employed in real-world applications, such as intelligent housing systems, law enforcement and real-time online meeting. Modern speaker verification technologies are mainly based on *deep speaker embedding* approach, which maps a piece of variable length speech to a fixed-dimension embedding through deep neural networks. The speaker embeddings extracted by ASV also serve as a key component for speaker diarization, voice conversion / cloning and speech recognition.

Convolution neural networks (CNNs) based models have achieved a remarkable success for ASV. X-vector [4] firstly employs the time delay neural networks (TDNNs) to maps variable-length utterances to fixed-dimensional embeddings.

Later, equipped with residual connections, ResNet-based [7, 8] r-vector system and its variantions [5, 9, 10] become capable of training deeper networks and have shown a outstanding results. Recently, based on blocks of TDNNs and squeeze and excitation (SE) [11] layers unified with Res2Block [12], the ECAPA-TDNN [6] and its subsequent efforts [13, 14] achieve a significant breakthrough and deliver the equal error rates less than 1% in VoxCeleb1-O benchmark.

Despite the great success, CNN still has its limitations. It mainly focuses on local spatial modeling, but lacks of global context fusion. CNNs-based models can not handle the long-range dependencies very well. To overcome this issue, Transformer [15] and its variations [16–18] have become a prevalent architecture in many sequence processing tasks. Transformers are good at modeling long-range global context and facilitate efficient parallel training. However, many existing studies such as UniSpeech [19] and WavLM [20, 21], indicate that without complicated pre-training procedures and large parameters, Transformer can hardly achieve a satisfactory performance in ASV. Recently, Convolution-augmented Transformer (Conformer) [22], a combination of CNNs and Transformers, becomes a promising candidate for advancing speech processing performance. Conformer inserts convolution modules into Transformer to increase the local information modeling. It firstly achieves an outstanding results in end-to-end speech recognition, and is later adopted in speech enhancement [23] and speech separation [24] with remarkable performance.

Inspired by these recent progresses, we proposed MFA-Conformer, an easy-to-implement and effective backbone for speaker embedding extraction. Firstly, the input acoustic feature is processed by a convolution subsampling layer to decrease the computational cost. Secondly, we adopt Conformer blocks which combine Transformers and convolution neural networks to capture the global and local features effectively. Finally, we concatenate the output features from all Conformer blocks to aggregate the multi-scale representations before final pooling. Our contributions can be summarized as follow:

- To our best knowledge, this study is the first to adopt Conformer blocks with elaborated modification for ASV. This work reveals that the Transformer-based model can achieve a remarkable performance in ASV without any complicated pre-training procedures or large network parameters.
- The proposed MFA-Conformer significantly outperforms the popular CNNs-based ECAPA-TDNN baseline systems in both inference speed and accuracy. Other than that, it achieves state-of-the-art results on SITW [25] benchmark.
- We conduct numerous experiments and the results demon-

---

† Corresponding author.
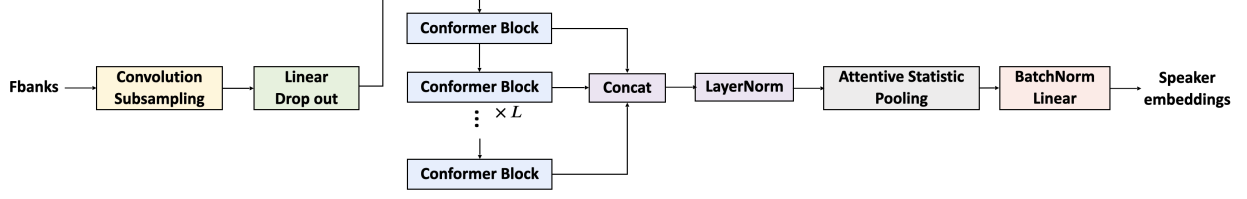[1]`https://github.com/zyzisyz/mfa_conformer`

Figure 1: *The overall architecture of Multi-scale Feature Aggregation Conformer (MFA-Conformer)*

strate the combination of local and global modeling can lead to the robust speaker embedding extraction. Especially in the real-world where the speech lengths are different, the MFA-Conformer shows robust performance as mentioned in Section 4.2

2. MFA-Conformer

In this section, we illustrate the fundamental components of the proposed Multi-scale Feature Aggregation Conformer (MFA-Conformer) system. The overall architecture is shown in Fig.1.

2.1. The Conformer block

Local features and global dependencies are both crucial for speaker representation learning. Local features, including the pitch, intonation style and pronunciation pattern, are essential to represent the speaker characteristics. And the good capacity of global context modeling to capture the long-range dependencies of variable-length speech will lead to robust speaker embedding extraction. Convolution neural networks are good at extracting local features but is weak at capturing global representations. Self-attention from Transformer can capture long-range global context dependencies but is lack of local details.

In order to both model the global and local features more directly and efficiently, we employ the Conformer block [22], a combination of CNNs and Transformers, to better capture the speaker characteristics. Multi-head self-attention (MHSA) and convolution modules are the key components for a Conformer block. The MHSA in the Conformer employs the relative positional encoding scheme proposed in Transformer-XL [16]. It encodes inputs according to the relative position deviation and takes account of both the global content offset and global position offset. Therefore, the relative positional encoding MHSA module is more robust for the utterances with variable input lengths. The convolution module after the MHSA consists of Pointwise convolution, 1D Depthwise convolution and BatchNorm after the convolution layer contributing to training deep models easier.

The structure of the Conformer block [22] is different from the Transformer block [15]. It contains two Macaron-like feed forward modules (FNN) with half residual connections sandwiching the MHSA and convolution modules (Conv). Mathematically, for the input feature h_{i-1} to the i Conformer block, the output feature h_i is calculated by:

$$
\begin{aligned}
\widetilde{h}_i &= h_{i-1} + \frac{1}{2}\mathrm{FNN}(h_{i-1}) \\
h_i' &= \widetilde{h}_i + \mathrm{MHSA}(\widetilde{h}_i) \\
h_i'' &= h_i' + \mathrm{Conv}(h_i') \\
h_i &= \mathrm{LayerNorm}(h_i'' + \frac{1}{2}\mathrm{FNN}(h_i''))
\end{aligned}
\tag{1}
$$

Note that $h_{i-1} \in \mathbb{R}^{d \times T}$ and $h_i \in \mathbb{R}^{d \times T}$, where d denotes the Conformer encoder dimension and T denotes the frame length.

2.2. MFA with attentive statistics pooling

Previous studies [26,27] indicate that the low-level feature maps can also contribute towards the accurate speaker embedding extraction. Based on this experience, in ECAPA-TDNN system, the output feature maps from all SE-Res2blocks are aggregated before the final pooling layer, and this aggregation leads to an obvious performance improvement. Likewise, we concatenate the output feature maps from each Conformer block and then feed them into a LayerNorm layer:

$$
\begin{aligned}
H' &= \mathrm{Concat}(h_1, h_2, ..., h_L) \\
H &= \mathrm{LayerNorm}(H')
\end{aligned}
\tag{2}
$$

where $H' \in \mathbb{R}^{D \times T}$ and $H = [H_1, H_2, ..., H_T] \in \mathbb{R}^{D \times T}$. L denotes the number of Conformer blocks and $D = d \times L$.

Furthermore, we adopt the attentive statistics pooling [6,28] to capture the importance of each frame and extract more robust speaker embedding. Specifically, for a frame-level feature H_t at time step t, we firstly calculate scalar score e_t and normalized score α_t as:

$$
\begin{aligned}
e_t &= \boldsymbol{v}^T f(\boldsymbol{W} H_t + \boldsymbol{b}) + k \\
\alpha_t &= \frac{\exp(e_t)}{\sum_{\tau=1}^{T} \exp(e_\tau)}
\end{aligned}
\tag{3}
$$

where $\boldsymbol{W} \in \mathbb{R}^{D \times D}$, $\boldsymbol{b} \in \mathbb{R}^{D \times 1}$, $\boldsymbol{v} \in \mathbb{R}^{D \times 1}$ and k are the trainable parameters for attention. $f(\cdot)$ denotes the Tanh activation function. After that, the normalized score α_t is adopted as the weight to calculate the weighted mean vector $\widetilde{\boldsymbol{\mu}}$ and weighted standard deviation $\widetilde{\boldsymbol{\sigma}}$, which are formulated as:

$$
\begin{aligned}
\widetilde{\boldsymbol{\mu}} &= \sum_{t=1}^{T} \alpha_t H_t \\
\widetilde{\boldsymbol{\sigma}} &= \sqrt{\sum_{t=1}^{T} \alpha_t H_t \odot H_t - \mu \odot \mu}
\end{aligned}
\tag{4}
$$

where $\mu = \frac{1}{T}\sum_{\tau=1}^{T} H_\tau$ and $\odot$ denotes the Hadamard product. The output of the pooling layer is given by concatenating the vectors of the weighted mean $\widetilde{\boldsymbol{\mu}}$ and weighted standard deviation $\widetilde{\boldsymbol{\sigma}}$.

Finally, the speaker embedding is extracted from a high dimension vector to a low dimension vector with BatchNorm using the fully-connected linear layer.

3. Experimental Setup

In this section, we present datasets, implementation details and evaluation protocols. In industrial applications scenario, such

as video processing & analysis, the speech lengths may vary from less than 5 seconds to more than 30 seconds. Therefore, VoxCeleb1-O and SITW benchmarks are adopted to illustrate the advantages of MFA-Conformer in different utterance duration scenarios.

3.1. Dataset

VoxCeleb1&2 [29, 30] and SITW [25] are used in our experiments. VoxCeleb is an audio-visual dataset consisting of 2,000+ hours short clips of human speech, extracted from interview videos on YouTube. SITW is a widely-used standard evaluation dataset in real-world conditions, consisting of 299 speakers including two testing trials (SITW.Dev and SITW.Eval). For model training, we use the development set of VoxCeleb1&2, which contain 1,240,000+ utterances from 7,205 speakers.

To better illustrate the advantages of MFA-Conformer in different utterance duration conditions, we adopt 3 trials including Voxceleb1-O, SITW.Dev and SITW.Eval for recognition performance evaluation. Note that Voxceleb1-O is the test part of Voxceleb1 and the major durations of testing utterances are 5-8s, which can be regarded as the short-duration utterance scenario. As for SITW, the major durations are about 30-40s, therefore it can be regarded as the long-duration scenario.

3.2. Network configurations

The speaker embedding dimension of all systems is 192. For fair comparisons, we also re-implement the r-vector system proposed in [8] and the ECAPA-TDNN system proposed in [6] as the baselines. Other configurations are presented below:

ResNet34. The first baseline is the ResNet-based r-vector system, which contains four residual blocks with different channels. We set the channels of residual blocks as {64, 128, 256, 512}. The total number of learnable parameters is 23.2M.

ECAPA-TDNN. The second baseline is ECAPA-TDNN, which contains three carefully designed SE-Res2Blocks. We set the channels of SE-Res2Blocks as {1024, 1024, 1024}. The total number of learnable parameters is 20.8M.

MFA-Conformer. The proposed MFA-Conformer, whose structure follows the practical experience in end-to-end speech recognition. Specifically, for multi-headed self-attention module, we set the encoder dimension as 256 and set the number of attention heads as 4; for convolution module, we set the kernel size to 15; for feed forward module, we set linear hidden units as 2048. We adopt 6 Conformer blocks with different subsampling rates. The total number of learnable parameters is about 19.7M-20.5M, which is closed to the above baseline systems.

3.3. Implementation details

We use Pytorch [31] framework to implement the proposed MFA-Conformer and the baseline systems[2]. A fixed length 3-second segments are extracted randomly from each utterance. The input features are 80-dimensional Fbanks with a window length of 25 ms and a frame-shift of 10 ms. No voice activity detection or augmentation are performed. All models are trained using additive margin Softmax (AM-Softmax) loss [33] with a margin of 0.2 and a scaling factor of 30. We use the

---

[2]Original code of the baseline systems is not available. For a fair comparison, all the models are trained in the same framework. There may be a few mismatches between the re-implemented baseline systems and reference papers [6, 8]. We borrow code from WeNet toolkit [32] and the training details can be found in the accompanying open-sourced code.

Adam optimizer with an initial learning rate of 0.001 and decrease the learning rate by 50% every 4 epoch. We also set the weight decay as 1e-7 to avoid overfitting and perform a linear warmup at the first 2k steps. The batch size is 200.

3.4. System evaluation

We use cosine distance with adaptive s-norm [34] for scoring. Then we report the Equal Error Rate (EER) and minimum Detection Cost Function (minDCF) with $P_{target} = 0.01$ and $C_{FA} = C_{Miss} = 1$ for performance evaluation. Furthermore, we calculate the real-time factor (RTF) on an Intel(R) Xeon(R) Silver 4210R CPU (2.40GHz) to evaluate the inference speed.

4. Experimental Results

4.1. Results on VoxCeleb test and SITW

In this section, we present the performance of the proposed MFA-Conformer with different subsampling rates, as well as the performance of the two baseline systems ResNet34 and ECAPA-TDNN. Table 1 reports the equal error rate (EER) and minimum Detection Cost Function (minDCF) together with the number of model parameters and real time factor (RTF). Firstly, it can be observed from the first and second lines that comparing with ResNet34, the ECAPA-TDNN system achieves an obvious advantage in recognition performance, yet the RTF is unsatisfied. Secondly, from the results of the proposed MFA-Conformer with different subsampling rates, we find that MFA-Conformer (1/1), MFA-Conformer(1/2) and MFA-Conformer (1/4) achieve promising results. Compared with the popular ECAPA-TDNN systems, the MFA-Conformer(1/2) obtains 21% relative improvement in EER and 32% relative improvement in inference speed (RTF). Thirdly, as described in Section 3.1, VoxCeleb1-O is a short-duration test scenario and SITW is a long-duration test scenario. We can find that the MFA-Conformer is able to obtain more competitive results than CNNs-based systems in long-duration utterance scenarios. This also indicates that MFA-Conformer can better handle long-range sequences and extract a more reliable speaker embedding for long utterances.

4.2. Impacts of global feature modeling

In the real-world applications, such as video processing, real-time online meeting, the utterance lengths may vary from less than 5 seconds to more than 30 seconds. Extracting robust global features for utterances with different lengths is important for speaker verification. The Multi-head self-attention (MHSA) is the key component to make Transformer or Conformer unique from the widely used CNNs-based models. In this section, we investigate the impacts of the global interactions modeling of MHSA by comparing the system performances with different utterance durations. We randomly split the test utterances of SITW according to different duration ranges to generate new trials, and report the EERs of the new trials in Fig.2. we can observe that when the test utterance duration becomes longer, the MFA-Conformer achieves larger relative improvements. It further indicates that MFA-Conformer is better at modeling long-range, global context information.

4.3. Impacts of local feature modeling

The Conformer inserts an additional Convolution module into the Transformer to better capture the local information. In this section, we study the impacts of the local feature model-

Table 1: *Performance overview of all systems on VoxCeleb1-O (short-duration), SITW.Dev & SITW.Eval (long-duration)*

Model	# Parameters	RTF	VoxCeleb1-O		SITW.Dev		SITW.Eval	
			EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
ResNet34	23.2M	0.0088	1.03	0.112	2.34	0.209	2.54	0.226
ECAPA-TDNN	20.8M	0.0180	0.82	0.112	1.91	0.179	2.22	0.192
MFA-Conformer (1/1)	20.5M	0.0203	0.83	0.102	1.51	0.159	1.78	0.172
MFA-Conformer (1/2)	20.5M	0.0121	**0.64**	**0.081**	**1.29**	**0.137**	**1.63**	**0.153**
MFA-Conformer (1/4)	19.8M	0.0102	0.83	0.118	1.88	0.160	1.94	0.178
MFA-Conformer (1/6)	20.4M	0.0093	1.22	0.142	1.91	0.200	2.46	0.222
MFA-Conformer (1/8)	19.7M	0.0089	1.48	0.182	2.88	0.240	2.79	0.261

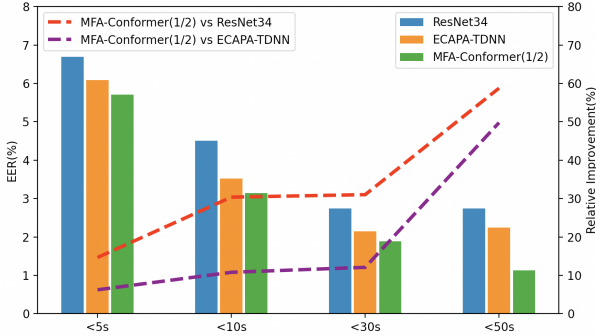*MFA-Conformer (1/2) means the convolution subsampling rate is 1/2.

Figure 2: *Performance of MFA-Conformer (1/2) and two baselines with different utterance durations. The bars denote the EERs, and the dotted lines denote the relative improvement of MFA-Conformer (1/2) over the two baseline.*

ing by comparing the performances between Transformer and Conformer blocks. We replace the Conformer blocks by Transformer blocks and make a group of experiments with different number of blocks L in SITW test set. As shown in Table 2, it can be clearly observed that Conformer blocks have remarkable advantages over Transformer blocks. MFA-Transformer can hardly obtain a satisfactory performance even increasing the number of the blocks. This indicates that the local spatial modeling by convolution module plays a critical role in accurate speaker embedding extraction. And the results show that setting L to 6 performs better than the rests.

Table 2: *MFA-Conformer v.s. MFA-Transformer on SITW*

Block	L	SITW.Dev		SITW.Eval	
		EER(%)	minDCF	EER(%)	minDCF
	1	4.78	0.357	4.96	0.413
	3	3.47	0.272	3.31	0.303
Transformer Block	6	2.50	0.237	3.01	0.246
	9	2.56	0.221	2.48	0.232
	12	2.61	0.224	2.65	0.238
	1	3.38	0.280	3.73	0.308
	3	2.15	0.186	2.14	0.195
Conformer Block	6	**1.29**	**0.137**	**1.63**	0.153
	9	1.45	0.141	1.76	0.158
	12	1.73	0.150	1.69	**0.149**

4.4. Ablation Studies

In the final section, we remove the individual components introduced in Section 2 to study the effect of each component

contributing to performance improvements. Due to the space limitation, we only present the results on VoxCeleb1-O shown in Table 3, and the results in other sets attain the same trend. The results in the first line are the MFA-Conformer with 1/2 subsampling rate. In the second line, we remove relative positional encoding scheme; in the third line, we remove the Macaron-style feed forward network and only keep the single feed-forward network after MHSA; in the fourth line, we discard the multi-scale feature aggregation strategy and only use the output representations from the last Conformer block; in the last line, we remove the convolution module to further measure the impact of the local feature modeling. It can be observed that multi-scale feature aggregation and convolution module play the most critical roles in achieving the promising performance Aggregation of the outputs from all blocks brings 48.3% relative improvement in EER. And convolution module leads to 54.9% relative improvement in EER.

Table 3: *Ablation study of MFA-Conformer on VoxCeleb1-O.*

	EER(%)	minDCF
MFA-Conformer (1/2)	**0.64**	**0.081**
w/o Relative PE	0.77	0.086
w/o Macaron FFN	0.84	0.085
w/o MFA	1.24	0.150
w/o Conv	1.42	0.147

*w/o is without.

5. Conclusions

This paper proposes MFA-Conformer, a novel backbone for automatic speaker verification. MFA-Conformer could be an ideal backbone for real industry speaker recognition scenarios. It significantly outperforms the popular ECAPA-TDNN systems in both recognition performance and inference speed. And it can extract more robust embeddings when the utterances are with different durations. Our ablation study shows the combination of local and global feature modeling can lead to the robust speaker embedding extraction, this can provide inspiration for the future ASV system design and acceleration. In the future work, we will extend the MFA-Conformer for streaming speaker recognition scenarios.

6. Acknowledgements

7. References

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

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

[3] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

[4] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[5] 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," *arXiv preprint arXiv:2003.11982*, 2020.

[6] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[8] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, "But system description to voxceleb speaker recognition challenge 2019," *arXiv preprint arXiv:1910.12592*, 2019.

[9] J.-w. Jung, H.-s. Heo, j.-h. Kim, H.-j. Shim, and H.-j. Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," *Proc. Interspeech 2019*, pp. 1268–1272, 2019.

[10] T. Zhou, Y. Zhao, and J. Wu, "Resnext and res2net structures for speaker verification," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 301–307.

[11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[12] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 652–662, 2019.

[13] J. Thienpondt, B. Desplanques, and K. Demuynck, "Integrating frequency translational invariance in tdnns and frequency positional information in 2d resnets to enhance speaker verification," *arXiv preprint arXiv:2104.02370*, 2021.

[14] T. Liu, R. K. Das, K. A. Lee, and H. Li, "Mfa: Tdnn with multi-scale frequency-channel attention for text-independent speaker verification with short utterances," *arXiv preprint arXiv:2202.01624*, 2022.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[16] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.

[17] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[18] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proceedings of AAAI*, 2021.

[19] C. Wang, Y. Wu, Y. Qian, K. Kumatani, S. Liu, F. Wei, M. Zeng, and X. Huang, "Unispeech: Unified speech representation learning with labeled and unlabeled data," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 937–10 947.

[20] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *arXiv preprint arXiv:2110.13900*, 2021.

[21] Z. Chen, S. Chen, Y. Wu, Y. Qian, C. Wang, S. Liu, Y. Qian, and M. Zeng, "Large-scale self-supervised speech representation learning for automatic speaker verification," *arXiv preprint arXiv:2110.05777*, 2021.

[22] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[23] Y. Koizumi, S. Karita, S. Wisdom, H. Erdogan, J. R. Hershey, L. Jones, and M. Bacchiani, "Df-conformer: Integrated architecture of conv-tasnet and conformer using linear complexity self-attention for speech enhancement," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2021, pp. 161–165.

[24] S. Chen, Y. Wu, Z. Chen, J. Wu, J. Li, T. Yoshioka, C. Wang, S. Liu, and M. Zhou, "Continuous speech separation with conformer," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5749–5753.

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

[26] Z. Gao, Y. Song, I. V. McLoughlin, P. Li, Y. Jiang, and L.-R. Dai, "Improving aggregation and loss function for better embedding learning in end-to-end speaker verification system." in *INTERSPEECH*, 2019, pp. 361–365.

[27] Y. Tang, G. Ding, J. Huang, X. He, and B. Zhou, "Deep speaker embedding learning with multi-level pooling for text-independent speaker verification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6116–6120.

[28] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *arXiv preprint arXiv:1803.10963*, 2018.

[29] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[30] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[31] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.

[32] Z. Yao, D. Wu, X. Wang, B. Zhang, F. Yu, C. Yang, Z. Peng, X. Chen, L. Xie, and X. Lei, "Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," in *Proc. Interspeech*. Brno, Czech Republic: IEEE, 2021.

[33] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[34] P. Matejka, O. Novotný, O. Plchot, L. Burget, M. D. Sánchez, and J. Cernocký, "Analysis of score normalization in multilingual speaker recognition." in *INTERSPEECH*, 2017, pp. 1567–1571.