

IMPROVING TRANSFORMER-BASED NETWORKS WITH LOCALITY FOR AUTOMATIC SPEAKER VERIFICATION

Mufan Sang^{1*}, Yong Zhao², Gang Liu², John H.L. Hansen¹, Jian Wu²

¹The University of Texas at Dallas, TX, USA

²Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA

{mufan.sang, john.hansen}@utdallas.edu, {yonzhao, ganli, jianwu}@microsoft.com

ABSTRACT

Recently, Transformer-based architectures have been explored for speaker embedding extraction. Although the Transformer employs the self-attention mechanism to efficiently model the global interaction between token embeddings, it is inadequate for capturing short-range local context, which is essential for the accurate extraction of speaker information. In this study, we enhance the Transformer with the locality modeling in two directions. First, we propose the Locality-Enhanced Conformer (LE-Conformer) by introducing depth-wise convolution and channel-wise attention into the Conformer blocks. Second, we present the Speaker Swin Transformer (SST) by adapting the Swin Transformer, originally proposed for vision tasks, into speaker embedding network. We evaluate the proposed approaches on the VoxCeleb datasets and a large-scale Microsoft internal multilingual (MS-internal) dataset. The proposed models achieve 0.75% EER on VoxCeleb 1 test set, outperforming the previously proposed Transformer-based models and CNN-based models, such as ResNet34 and ECAPA-TDNN. When trained on the MS-internal dataset, the proposed models achieve promising results with 14.6% relative reduction in EER over the Res2Net50 model.

Index Terms— Speaker verification, self-attention, Transformer, Conformer, Swin Transformer

1. INTRODUCTION

In the past few years, we have seen the fast development of deep neural network (DNN) based speaker verification (SV). Various approaches for speaker verification have been proposed with different DNN architectures [1, 2, 3], novel loss functions [4, 5, 6], and frameworks for domain mismatch [7, 8, 9]. In [10, 11, 12], researchers further studied semi-supervised and self-supervised SV systems using partial-labeled or unlabeled data.

For speaker embedding extraction, x-vector [1] has proven its success by utilizing the 1D convolutional neural network (CNN) based time delay neural network (TDNN). Moreover, 2D CNNs (i.e. ResNet-based architectures) are also successfully adopted to the SV task and obtain remarkable performance [13, 5]. ECAPA-TDNN [3] was proposed to further enhance the TDNN-based architecture and achieved a competitive performance with ResNet. These predominant CNN-based models take advantage of strong ability of capturing local speaker patterns from speech features. To further improve the performance of CNN-based speaker embedding networks, some attention mechanisms were integrated to the speaker embedding extractor [14, 15] or the pooling layer [13, 16]. Convolution

layer allows CNNs to model the local dependencies well, but it lacks a mechanism to capture speaker information globally. There have been attempts to explore using Transformer to replace CNNs for speaker embedding extraction [17, 18, 19]. However, without large-scale pre-training, Transformer-based speaker embedding networks can hardly achieve competitive performance as CNNs for speaker verification. This is primarily due to the lack of certain desirable properties inherently built into the CNN architecture such as locality.

To enable Transformers to capture global and local context collectively, in this study, we introduce locality mechanisms to Transformer in two different directions: (1) integrating depth-wise convolution and channel-wise attention into Transformer block, (2) introducing a hierarchical Transformer architecture with shifted local window self-attention, inspired by Swin Transformer [20]. Consequently, we propose two Transformer-based speaker embedding networks: Locality-Enhanced Conformer (LE-Conformer) and Speaker Swin Transformer (SST). For Locality-Enhanced Conformer, we propose adding depth-wise convolution and channel-wise attention into the feed-forward network (FFN) of Conformer [21]. Furthermore, we study the effective way to aggregate the output features from all the Conformer blocks to enhance the frame-level speaker embedding. For Speaker Swin Transformer, we introduce a hierarchical structure that produces multi-scale output feature maps with local window self-attention. Experimental results on the VoxCeleb datasets indicate that the proposed LE-Conformer and SST significantly outperform previous Transformer-based models and strong ResNet and ECAPA-TDNN baseline systems. Moreover, when trained on a larger-scale MS-internal multilingual dataset, the proposed systems outperform Res2Net50 by a large margin, producing more robust and competitive speaker embeddings. The primary contributions of this paper can be summarized as follow: (1) We propose an effective locality mechanism for Conformer to enhance the ability of local information aggregation. (2) We propose the Speaker Swin Transformer which generates multi-scale output feature maps with shifted local window self-attention. (3) We conduct comprehensive experiments to demonstrate the effectiveness of the proposed Transformer-based networks with locality mechanisms.

2. LOCALITY-ENHANCED CONFORMER

Self-attention is the key component of Transformer. It enables Transformer to have a strong ability to model global interaction between speech frames. However, global self-attention does not have sufficient ability to capture local information which is essential for speaker embedding. Therefore, we introduce locality mechanisms

*Work performed while Mufan Sang was an intern at Microsoft.

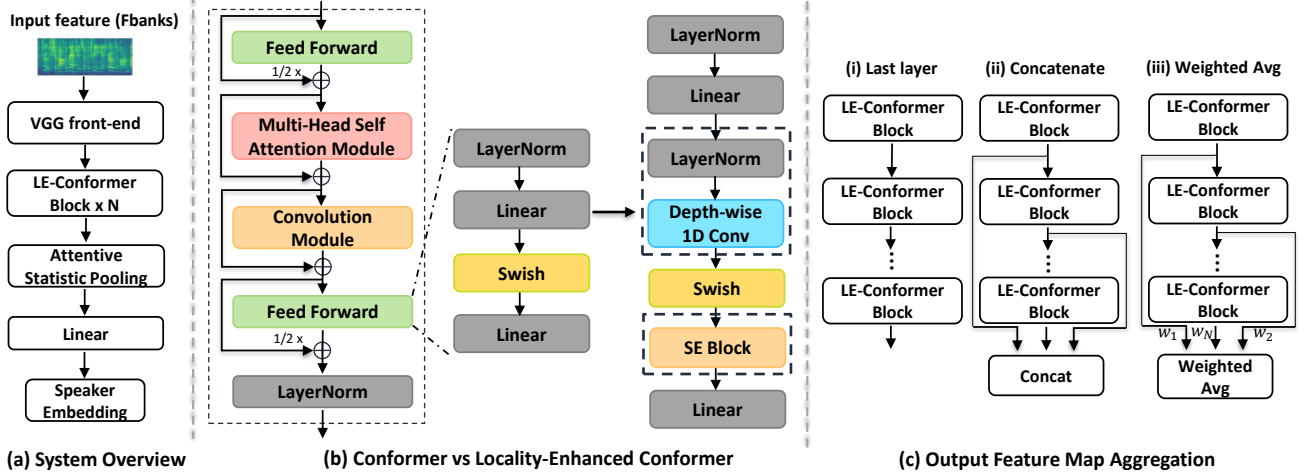


Fig. 1: The architecture of Locality-Enhanced Conformer for speaker embedding.

to enhance Transformer in modeling local dependencies. In this section, we present Locality-Enhanced Conformer (LE-Conformer), which is built upon the Conformer [21] architecture. We incorporate additional convolution, channel-wise attention, and intermediate output feature map aggregation into the Conformer network. The system architecture is illustrated in Fig. 1(a). It comprises (i) a VGG front-end to subsample the input feature, (ii) a number of locality-enhanced Conformer blocks to extract frame-level speaker embedding, (iii) a pooling layer to generate utterance-level embedding, (iv) a linear layer to extract the final speaker embedding.

2.1. Conformer Encoder Block

As a state-of-the-art model in ASR, Conformer [21] combines convolution and self-attention in the Transformer block to enhance its capability of capturing local information. The architecture of Conformer block is shown in the left side of Fig. 1(b). It consists of multi-head self-attention (MSA) module and convolution module sandwiching by two Macron-style feed-forward networks (FFN). Assuming z_i as the input for the i -th Conformer block, the output of this block z_{i+1} is computed as

$$\begin{aligned} \tilde{z}_i &= z_i + \frac{1}{2} \text{FFN}(z_i) \\ z'_i &= \tilde{z}_i + \text{MSA}(\tilde{z}_i) \\ z''_i &= z'_i + \text{Conv}(z'_i) \\ z_{i+1} &= \text{LayerNorm}\left(z''_i + \frac{1}{2} \text{FFN}(z''_i)\right) \end{aligned} \quad (1)$$

where FFN denotes the feed-forward network, MSA denotes the multi-head self-attention, and Conv denotes the convolution module.

2.2. Introducing Locality and Channel-wise Attention

In each block, adding convolutional layers after MSA sequentially helps Conformer capture local relationships. Regarding the FFN module, hidden dimension of the latent feature is expanded between the two linear layers. We consider that depth-wise convolution could be a filter for the latent representation and introduce additional locality to Transformer. To achieve this goal, we propose an effective strategy to integrate depth-wise convolution and a channel-wise attention mechanism into the feed-forward network.

As shown in the right side of Fig. 1(b), LayerNorm and a 1D depth-wise convolution layer are added after the first linear layer of FFN. The depth-wise convolution provides information interaction among adjacent frames, and captures local continuity of input feature maps. Integrating Squeeze-and-Excitation (SE) block [22] after convolution can further improve representation ability by modeling the inter-dependencies between channels of feature maps and re-calibrating each channel.

2.3. Aggregating Output Feature Maps of LE-Conformer Blocks

For the frame-level speaker embedding extractor, the output of the last layer is often used as the input for the pooling layer. However, previous studies [23, 24] indicated that feature maps from lower layers can also contribute to robust speaker embeddings. Similar to [25], we apply two strategies to utilize low-layer feature maps. As illustrated in Fig. 1(c), the first strategy is to concatenate the output feature maps from all LE-Conformer blocks along the channel dimension. Secondly, we conduct weighted average of the output feature maps from all LE-Conformer blocks with learnable weights.

Finally, the enhanced frame-level speaker embedding is processed by the pooling layer to generate utterance-level speaker embedding.

3. SPEAKER SWIN TRANSFORMER

Generally, Transformer computes the self-attention globally which contributes to a great long-range dependencies modeling capability. However, speech features can be much longer than text sentences. Thus, Transformers usually incur high computational and memory costs for speech tasks. Moreover, with convolution operation, ResNet-based networks bring locality, inductive bias, and multi-scale feature maps which contribute to their successes on SV tasks. On the contrary, Transformer's output feature maps are single-scale among all blocks. With a hierarchical structure and shifted window self-attention, Swin Transformer [20] has achieved state-of-the-art performance on multiple vision tasks. Therefore, we bring the advantages of Swin Transformer to speaker verification, proposing the Speaker Swin Transformer (SST), which generates multi-scale output features and computes self-attention with shifted local windows. The architecture leads to linear computational complexity to the length of utterance. The overview of the Speaker Swin Transformer encoder is presented in Fig. 2.

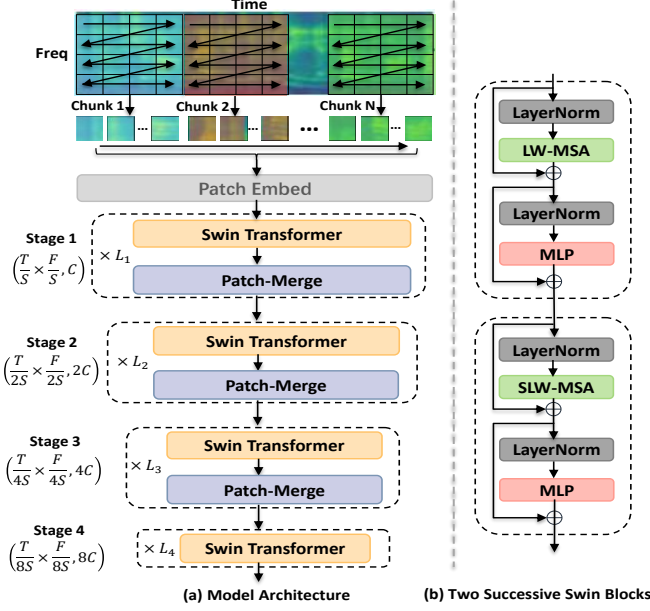


Fig. 2: The architecture of Speaker Swin Transformer.

3.1. Encoding Input Features by Overlapping Patch Embedding

Different from Conformer, Speaker Swin Transformer processes input speech features patch-by-patch instead of frame-by-frame. Considering the characteristics of speech feature, using patch embedding as input token allows model to learn both the temporal and frequency structure. As shown in Fig. 2(a), to better capture local continuity of input feature, the input feature is firstly split into a sequence of $P \times P$ ($P = 7$) patches with overlap by half of its area. Each patch is flattened and projected into a C -dimensional ($C = 96$) 1D patch embedding with a linear patch embedding layer. We utilize a convolution layer with 7×7 kernel size, a stride of 4 ($S = 4$), and zero padding of 3 for patch embedding. Accordingly, the output size of patch embedding is $(\frac{T}{S} \times \frac{F}{S}, C)$, where T and F represent time and frequency domain of input feature. Although Fbanks and images have a similar format with 2D shape, the height and width of Fbanks contain different information which represents frequency and temporal dimension. Inspired by [26], to better model the dependencies among frequencies for nearby frames, we first split the whole Fbanks into chunks along temporal dimension. Then, patches are split within each chunk following the order shown in Fig. 2(a), and all patches compose the sequence chunk-by-chunk. Then, the patch token sequence is processed by the linear patch embedding layer and sent to the following Transformer blocks.

3.2. Swin Transformer Block

In order to construct the hierarchical structure with local window self-attention, we adapt Swin Transformer block to speaker verification. Inside each Swin Transformer block, the local window self-attention is introduced to replace the conventional global self-attention for efficient modeling. Regarding self-attention computing, all the input patch tokens are evenly partitioned into non-overlapping local windows which contain $M \times M$ ($M = 5$) patches. Accordingly, self-attention is computed within each local window instead of among all patches globally. For an input speech feature with $f \times t$ patch tokens, the computational complexity of a global MSA is $\mathcal{O}(ftC^2 + (ft)^2C)$ and a local window MSA (LW-MSA)

is $\mathcal{O}(ftC^2 + M^2ftC)$ with the latent feature dimension C . It demonstrates that LW-MSA is much more efficient than global MSA with $M^2 \ll ft$ ($M = 5$) and its complexity grows linearly with ft instead of quadratically as global MSA. Therefore, it is able to substantially decrease the computation cost.

To enlarge the receptive field and model the connections across windows, shifted local window multi-head self-attention (SLW-MSA) is introduced in addition to the LW-MSA. This module adopts a windowing configuration that is shifted from that of the preceding layer, by shifted toward lower-right by $(\lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor)$ patches in consecutive Swin Transformer blocks. As shown in Fig. 2(b), the Transformer block with SLW-MSA introduces connections between neighboring non-overlapping windows from the previous block.

To form a hierarchical structure, a patch merging layer [20] is added at the end of each stage from stage 1 to 3. For the first patch merging layer, it concatenates the neighboring patches with group of 2×2 , and applies a linear layer to reduce the output dimension from $(\frac{T}{25} \times \frac{F}{25}, 4C)$ to $(\frac{T}{25} \times \frac{F}{25}, 2C)$. As illustrated in Fig. 2(a), the shape of patch tokens is reduced to $(\frac{T}{45} \times \frac{F}{45}, 4C)$ and $(\frac{T}{85} \times \frac{F}{85}, 8C)$ through stage 2 and 3, respectively. These stages allow the network to generate hierarchical representations as ResNet, so the memory cost is decreased exponentially through each stage. More importantly, the receptive field for local window self-attention grows larger as layer goes deeper. In summary, these designs contribute to efficient local and global relations modeling.

4. EXPERIMENTS

4.1. Datasets

VoxCeleb The SV systems are trained on development set of VoxCeleb1&2 [27, 28] and evaluated on VoxCeleb1 test set. The total duration of training data is around 2k hrs. We apply data augmentation including additive noise with MUSAN [29], reverberation with RIR [30] and speed perturbation (with speed factor 0.9 and 1.1). Specifically, the generated utterances by speed perturbation are regarded as from new speakers.

MS-internal This is a large-scale Microsoft internal multilingual dataset collected in controlled acoustic environments. It consists of around 26 million utterances from over 29k speakers in 48 language locales, about 4.6 seconds in length per utterance. The training set contains around 24 million utterances from over 27k speakers with total duration of 33k hrs, and the test set contains over 0.1 million utterances from around 1.7k speakers. There is no speaker overlapping between training and testing set.

4.2. Implementation Details

The input features are 80-dimensional log Mel-filterbanks with a frame-length of 25 ms and 10 ms shift, applied with mean normalization at the utterance level. The proposed Transformers serve as the backbone network. Attentive statistic pooling (ASP) [16] is used to generate utterance-level embeddings. The models are trained with additive margin softmax (AM-softmax) loss [6] with a margin of 0.2 and a scaling factor of 30. We also trained ResNet34, ECAPA-TDNN, and Res2Net50 for comparison.

Locality-Enhanced Conformer A segment of 2.0 seconds is randomly selected for each input utterance. The model consists of 6 Locality-Enhanced Conformer encoder blocks with 4 attention heads. For each block, we set the encoder dimension as 512, the kernel size of convolution module as 15, and the hidden unit size as 2048 for the FFN. We use the AdamW [31] optimizer with an initial learning rate of $3e-4$ and set the weight decay as $5e-2$. A

Table 1: Performance of all SV systems on VoxCeleb1. *Upper block: CNN-based models; Middle block: Transformer-based models.*

Systems	Corpus	EER(%)	minDCF
ResNet34 [2]	Vox1&2	1.06	0.084
ECAPA-TDNN [3]	Vox1&2	0.85	0.078
Res2Net50 [32]	Vox1&2	0.55	0.041
SAEP [18]	Vox2	5.44	—
Wang et al. [19]	Vox2	2.56	—
DT-SV [33]	Vox2	1.92	0.130
S-vector+PLDA [17]	Vox1&2	2.67	0.300
LE-Conformer (ours)	Vox1&2	0.75	0.055
SST (ours)	Vox1&2	1.34	0.104

linear warm-up is applied at the first 45k steps and the learning rate is adjusted based on cyclical annealing schedule in the range of $1e-8$ and $3e-4$. The batch size is 128 for each of 8 GPU cards.

Speaker Swin Transformer A segment of 3.2 seconds is randomly selected for each input utterance. The input feature with size 320×80 will be equally split into 2 chunks along time dimension with size 160×80 . We set patch size as 7×7 , and attention window size as 5×5 . The four network stages are designed with 2, 2, 6, 2 Swin Transformer blocks, respectively. We set the channel number of the hidden layer in the first stage $C = 96$. We utilize AdamW optimizer with an initial learning rate of $3e-2$ and set the weight decay as $5e-2$. A linear warm-up is applied at the first 130k steps, and the learning rate is adjusted based on cyclical annealing schedule in the range of $1e-5$ and $3e-2$. The batch size is 200 for each of 8 GPU cards.

We report the system performance using two evaluation metrics: Equal Error Rate (EER) and minimum Detection Cost Function (minDCF) with $p_{target} = 0.05$.

4.3. Evaluation on VoxCeleb Dataset

First, we evaluate the proposed LE-Conformer and SST on the VoxCeleb dataset. Table 1 compares the performance of our systems with CNN-based baseline systems ResNet34 [2], ECAPA-TDNN [3], Res2Net50 [32] and other Transformer models [17, 18, 19, 33] for speaker verification. The LE-Conformer and SST achieve 0.75% and 1.34% EER, respectively, which significantly outperform other Transformer-based systems. Compared to S-vector+PLDA, LE-Conformer and SST improve the performance with relative 71.9% and 49.8% reduction in EER. Moreover, the LE-Conformer and SST also outperform the Wang et al [19] with relative 70.7% and 47.7%, and DT-SV with 60.9% and 30.2% improvement in EER.

Compared to CNN-based networks, LE-Conformer outperforms ResNet34 and ECAPA-TDNN with relative 29.3% and 11.8% improvement in EER, respectively. It demonstrates that introducing locality mechanisms to the Transformer is beneficial for more accurate extraction of speaker embeddings.

We further conduct experiments to illustrate the effectiveness of the proposed speaker embedding networks, and how the local information helps to improve the performance of Transformer-based models. Table 2 shows the impact of changes to the LE-Conformer block and SST. For LE-Conformer, without adding SE block and depth-wise convolution into the FFN, the performance degrades with an increase of 25.3% and 23.6% relatively in EER and minDCF, respectively. Without concatenating the output feature maps from all

Table 2: Ablation study of the Locality-Enhanced Conformer and Speaker Swin Transformer. *Non-OPE: non-overlapping patch embedding.*

Systems	EER(%)	minDCF
LE-Conformer	0.75	0.055
No SE Block	0.87	0.060
No DW Conv	0.94	0.068
No Concat	1.00	0.070
Weighted avg	1.21	0.091
SST	1.34	0.104
Non-OPE	1.47	0.120

Table 3: Performance of the proposed systems on MS-internal dataset.

Systems	EER (%)	minDCF
Res2Net50	3.09	0.180
LE-Conformer	3.57	0.229
SST	2.64	0.168

blocks, the performance further degrades by 6.4% relative in EER. It is equivalent to Conformer after removing the SE block, DW conv, and concatenation. The results in Table 2 demonstrate that the performance of Transformer can be significantly improved by the integration of locality mechanisms. Moreover, the performance becomes even worse if output features are averaged with learnable weights before the pooling layer.

For SST, it is beneficial to use overlapping patch embedding (OPE), which yields 8.8% relative improvement in EER compared to non-overlapping patch embedding (non-OPE). The overlapping patch embedding can effectively enhance the capability of modeling local continuity of input features via overlapped sliding windows.

4.4. Evaluation on MS-internal Dataset

In this section, we investigate the performance of proposed models when trained on a large-scale Microsoft internal multilingual (MS-internal) dataset. We compare LE-Conformer and SST to Res2Net50, the best CNN-based system in Table 1. As illustrated in Table 3, Speaker Swin Transformer outperforms Res2Net50 by 14.6% relative improvement in EER. It demonstrates that the hierarchical transformer structure with shifted local window self-attention is capable of making full use of the massive training data and learning global and local information collectively, compared with CNN-based models.

5. CONCLUSIONS

In this paper, we propose two speaker embedding networks by incorporating locality mechanisms to Transformer. Firstly, we integrate depth-wise convolution and channel-wise attention into the Conformer blocks to enhance the ability of modeling local dependencies. Secondly, we introduce the Speaker Swin Transformer which processes input features at multiple scales with shifted local window self-attention. Experimental results demonstrate that our models significantly outperform previous Transformer-based models and CNN-based models, such as ResNet34 and ECAPA-TDNN. The proposed architectures enable the effective extraction of speaker embeddings, especially when trained on large amounts of data. We hope this work can provide inspirations for the ultimate design of Transformer-based speaker embedding networks.

6. REFERENCES

- [1] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *ICASSP*, 2018, pp. 5329–5333.
- [2] Hossein Zeinali, Shuai Wang, Anna Silnova, Pavel Matějka, and Oldřich Plchot, “But system description to voxceleb speaker recognition challenge 2019,” *arXiv preprint arXiv:1910.12592*, 2019.
- [3] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, “Ecapadnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” *arXiv preprint arXiv:2005.07143*, 2020.
- [4] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno, “Generalized end-to-end loss for speaker verification,” in *ICASSP*, 2018, pp. 4879–4883.
- [5] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee-Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han, “In defence of metric learning for speaker recognition,” *Proc. Interspeech 2020*, pp. 2977–2981, 2020.
- [6] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu, “Cosface: Large margin cosine loss for deep face recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5265–5274.
- [7] Mufan Sang, Wei Xia, and John H.L. Hansen, “Open-set short utterance forensic speaker verification using teacher-student network with explicit inductive bias,” *Proc. Interspeech 2020*, pp. 2262–2266, 2020.
- [8] Gautam Bhattacharya, Joao Monteiro, Jahangir Alam, and Patrick Kenny, “Generative adversarial speaker embedding networks for domain robust end-to-end speaker verification,” in *ICASSP*. IEEE, 2019, pp. 6226–6230.
- [9] Mufan Sang, Wei Xia, and John H.L. Hansen, “Deaan: Disentangled embedding and adversarial adaptation network for robust speaker representation learning,” in *ICASSP*. IEEE, 2021, pp. 6169–6173.
- [10] Nakamasa Inoue and Keita Goto, “Semi-supervised contrastive learning with generalized contrastive loss and its application to speaker recognition,” in *APSIPA ASC*. IEEE, 2020, pp. 1641–1646.
- [11] Mufan Sang, Haoqi Li, Fang Liu, Andrew O Arnold, and Li Wan, “Self-supervised speaker verification with simple siamese network and self-supervised regularization,” in *ICASSP*. IEEE, 2022, pp. 6127–6131.
- [12] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Zhuo Chen, Peidong Wang, Gang Liu, Jinyu Li, Jian Wu, Xiangzhan Yu, and Furu Wei, “Why does Self-Supervised Learning for Speech Recognition Benefit Speaker Recognition?,” in *Proc. Interspeech 2022*, 2022, pp. 3699–3703.
- [13] Weicheng Cai, Jinkun Chen, and Ming Li, “Exploring the encoding layer and loss function in end-to-end speaker and language recognition system,” in *Proc. Odyssey*, 2018, pp. 74–81.
- [14] Jianfeng Zhou, Tao Jiang, Zheng Li, Lin Li, and Qingyang Hong, “Deep speaker embedding extraction with channel-wise feature responses and additive supervision softmax loss function,” in *Interspeech*, 2019, pp. 2883–2887.
- [15] Mufan Sang and John H.L. Hansen, “Multi-Frequency Information Enhanced Channel Attention Module for Speaker Representation Learning,” in *Proc. Interspeech 2022*, 2022, pp. 321–325.
- [16] Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda, “Attentive statistics pooling for deep speaker embedding,” *arXiv preprint arXiv:1803.10963*, 2018.
- [17] Narla John Metilda Sagaya Mary, Srinivasan Umesh, and Sandesh Varadaraju Katta, “S-vectors and tesa: Speaker embeddings and a speaker authenticator based on transformer encoder,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 404–413, 2021.
- [18] Pooyan Safari, Miquel India, and Javier Hernando, “Self-attention encoding and pooling for speaker recognition,” *Proc. Interspeech 2020*, pp. 941–945, 2020.
- [19] Rui Wang, Junyi Ao, Long Zhou, Shujie Liu, Zhihua Wei, Tom Ko, Qing Li, and Yu Zhang, “Multi-view self-attention based transformer for speaker recognition,” in *ICASSP*. IEEE, 2022, pp. 6732–6736.
- [20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *ICCV*, 2021, pp. 10012–10022.
- [21] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., “Conformer: Convolution-augmented transformer for speech recognition,” *Proc. Interspeech 2020*, pp. 5036–5040, 2020.
- [22] Jie Hu, Li Shen, and Gang Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE CVPR*, 2018, pp. 7132–7141.
- [23] Zhifu Gao, Yan Song, Ian McLoughlin, Pengcheng Li, Yiheng Jiang, and Li-Rong Dai, “Improving aggregation and loss function for better embedding learning in end-to-end speaker verification system,” in *Proc. INTERSPEECH*, 2019, pp. 361–365.
- [24] Yun Tang, Guohong Ding, Jing Huang, Xiaodong He, and Bowen Zhou, “Deep speaker embedding learning with multi-level pooling for text-independent speaker verification,” in *ICASSP*. IEEE, 2019, pp. 6116–6120.
- [25] Yang Zhang, Zhiqiang Lv, Haibin Wu, Shanshan Zhang, Pengfei Hu, Zhiyong Wu, Hung-yi Lee, and Helen Meng, “Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification,” *arXiv preprint arXiv:2203.15249*, 2022.
- [26] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, “Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 646–650.
- [27] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, “Voxceleb: A large-scale speaker identification dataset,” *Proc. Interspeech 2017*, pp. 2616–2620, 2017.
- [28] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, “Voxceleb2: Deep speaker recognition,” *Proc. Interspeech 2018*, pp. 1086–1090, 2018.
- [29] David Snyder, Guoguo Chen, and Daniel Povey, “Musan: A music, speech, and noise corpus,” *arXiv preprint arXiv:1510.08484*, 2015.
- [30] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition,” in *ICASSP*. IEEE, 2017, pp. 5220–5224.
- [31] Ilya Loshchilov and Frank Hutter, “Decoupled weight decay regularization,” in *International Conference on Learning Representations*, 2018.
- [32] Tianyan Zhou, Yong Zhao, and Jian Wu, “Resnext and res2net structures for speaker verification,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 301–307.
- [33] Nan Zhang, Jianzong Wang, Zhenhou Hong, Chendong Zhao, Xiaoyang Qu, and Jing Xiao, “Dt-sv: A transformer-based time-domain approach for speaker verification,” *arXiv preprint arXiv:2205.13249*, 2022.