ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747172

MLP-SVNET : A MULTI-LAYER PERCEPTRONS BASED NETWORK FOR SPEAKER VERIFICATION

Bing Han, Zhengyang Chen, Bei Liu, Yanmin Qian[†]

MoE Key Lab of Artificial Intelligence, AI Institute
X-LANCE Lab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Convolution and self-attention based neural networks have both obtained excellent performance in automatic speaker verification. However, the convolution model often lacks the ability of long-term dependency modeling due to the limitation of receptive field, while the self-attention model is insufficient to model local information. To tackle this limitation, we propose a new multi-layer perceptrons based speaker verification network (MLP-SVNet) which can apply MLPs across temporal and frequency dimensions to capture the local and global information at the same time. The experimental results conducted on Voxceleb show that the proposed model is very competitive when compared to other systems based on convolution or self-attention. In addition, we demonstrate that MLP-SVNet based on multi-layer perceptrons can produce complementary embeddings, which can be fused with the state-of-the-art system to further improve the performance.

Index Terms— Multi-layer Perceptron, Speaker Verification, Speaker Embedding, Text-independent

1. INTRODUCTION

Speaker verification (SV) is a task that utilizes speech as the biometric feature to verify the speakers' identities. Recently, the end-to-end deep embedding learning methods have been broadly applied for SV task and obtained excellent performance [1, 2, 3, 4, 5]. Generally, these model architectures are composed of three deep neural network components including a frame-level feature extractor, an utterance-level representation aggregator, and a speaker classifier.

To further improve SV performance, many models with different network architectures have been proposed in recent studies. And most of the studies focus on the convolution structure, including time delay neural network (TDNN) [1], residual networks (ResNet) [2], Dual Path Network (DPN) [6], ECAPA-TDNN [7] and so on. Convolution Neural Networks (CNN) are known as shift invariant or space invariant artificial neural networks, based on the shared-weight architecture of the convolution kernels or filters that slide along input features [8]. This independence assumption from prior knowledge leads to an excellent ability to model local features. Benefiting from this, convolution based models have achieved excellent performance in SV tasks. However, due to the limitation of the receptive field, convolution lacks the ability to model long-term dependency. To solve this problem, [3] proposed a tandem self-attention encoder and pooling layer to obtain a discriminative speaker embedding, which is inspired by transformer's [9] high parallelization capabilities and strong performance on computer vision and natural language processing [10]. Although self-attention solves the problem of long-term information modeling, it is insufficient for capturing local information.

In this study, inspired by [11], we propose a new multi-layer perceptrons based speaker verification network (MLP-SVNet), which does not use any convolutions or self-attention mechanism and bases entirely on multi-layer perceptrons instead. It applies MLPs across either temporal or frequency for modeling the local and global information at the same time. Comparing with CNN or attention based models, 1) MLP-SVNet has less inductive bias and more trainable parameters which will bring better fitting ability. 2) MLP-SVNet across temporal and frequency dimensions can balance global and local information at the same time. 3) As a totally different architecture, MLP-SVNet can produce complementary speaker embeddings, which means the fusion with MLP-SVNet can lead to much more improvement than other systems.

The rest of the paper is organized as below: In Section 2, we introduce some related works about architecture design for speaker verification. Then, we present our proposed MLP-SVNet. Next, experimental results are presented and analyzed in Section 4. And finally, the conclusion is given in Section 5.

2. RELATED WORKS

2.1. X-vector and R-vector

The emergence of the x-vector [1] system marked that neural network based system completely outperforms the systems

[†]Yanmin Qian is the corresponding author

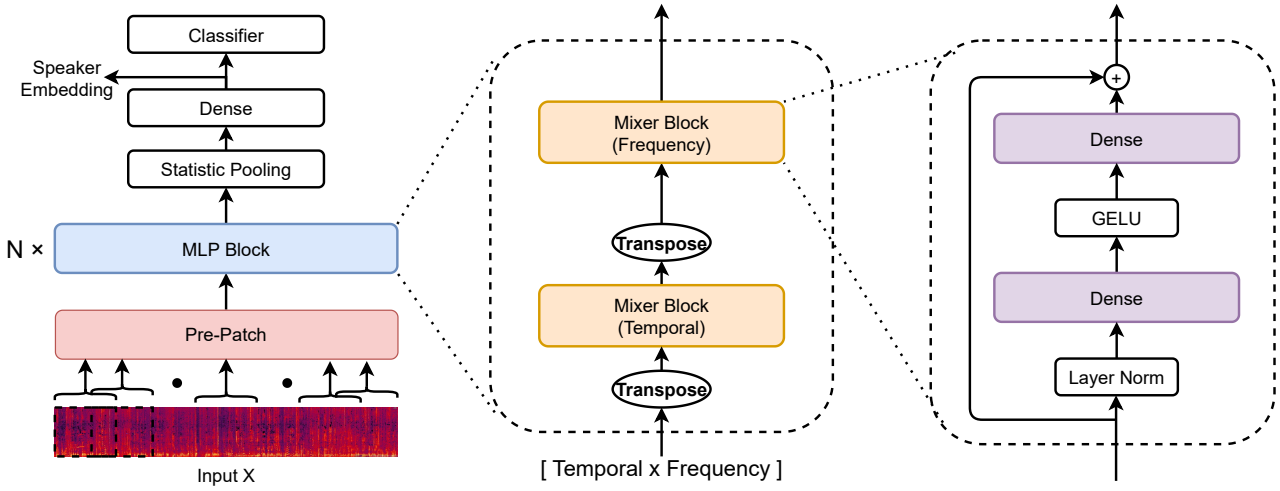

Fig. 1. The architecture of proposed MLP-SVNet. MLP-SVNet consists of pre-patch block, MLP blocks, statistic pooling layer, and a speaker classifier. Each MLP block contains one temporal Mixer block and one frequency Mixer block. And each Mixer block is composed of two fully-connected layers, layer norm, residual connection, and a GELU nonlinearity function.

based on i-vector [12]. X-vector has five time-delay layers to handle the input at the frame level, followed by a statistical pooling layer that computes the mean and standard deviation of the input sequence, which aggregates the frame-level input into a segment-level representation.

R-vector system is another convolution based network which is proposed in [2] and achieves superior performance in SV for its high-efficiency modeling complex data structure. Different from X-vector which only utilize convolution 1D to extract the feature, R-vector processes the features as a 2-dimensional signal before the statistic pooling layer. Finally, a fully connected layer transforms features into a fixed vector to represent the speaker.

2.2. S-vector

S-vector [3] is a new architecture, where the frame-level feature extractor is replaced with a transformer's [9] encoder which is based on self-attention. This mechanism is built on the dot product between frames, and allows the model to capture long-term speaker characteristics based on unrestricted context.

2.3. ECAPA-TDNN

ECAPA-TDNN is proposed in [7] and it's commonly acknowledged that it has been the state-of-the-art (SOTA) system nowadays. ECAPA-TDNN's architecture is an enhanced version of the conventional X-vector system. It integrates a Res2Net [13] module to enhance the central volume layer and constructs a hierarchical residual connection to handle multi-scale features. It also introduces 1-dimensional TDNN specific SE-blocks [14] which help the architecture to better model the channel interdependencies.

3. MLP-SVNET

MLP-SVNet presented in this paper can be mainly divided into four parts including a pre-patch layer, MLP blocks, a statistic pooling layer, and a dense classifier. In this section, we will give a detailed description of MLP-SVNet and an overview of the architecture is depicted in Figure 1.

3.1. Pre-Patch

D-vector [15] has demonstrated that stacking each training frame with its left and right context frames can provide better performance than a single frame. Inspired by this, we propose the pre-patch module in order to encode the neighbor information. As shown in Figure 1 (left), the input $\mathbf{X}$ is a feature map whose dimension is composed of temporal and frequency. Then, similar to the patching method in [10], we split the feature map into overlapped patches with a sliding window. Finally, all the patches are flattened to vectors and encoded into a sequence of fixed-dimension embeddings, which is used as the input to feed into MLP blocks in the following.

There are two patching methods we have proposed in this study: patch 1D and patch 2D. Patch 1D means that we split the feature map across temporal dimension and stack the frames with their neighbor contents. And the latter one treats the feature map as an image to split it into square patches across not only temporal but also frequency dimensions.

3.2. MLP Block

MLP-SVNet is mainly composed of multiple identical size MLP blocks, and each MLP block consists of two Mixer blocks to model the temporal and frequency information of the input $\mathbf{X} \in \mathbb{R}^{T \times F}$, where T and F are time and frequency

dimension respectively. As shown in Figure 1 (middle), the first Mixer block is the temporal Mixer block. It applies the dense transformation on the temporal dimension of the input. Because the temporal dimension is the column of the feature map, we add transpose operation before and after the temporal Mixer blocks for the convenience of implementation. The second one is the frequency Mixer block which applies the dense transformation on the frequency dimension in order to mix the frequency features.

For each Mixer block, it contains two dense layers, residual connection and an element-wise non-linearity activate function GELU [16] which is shown in Figure 1 (right). As described above, MLP blocks can be written as follows:

$$\mathbf{Y} = Mixer((Mixer(\mathbf{X}^\intercal))^\intercal) \tag{1}$$

where $^\intercal$ means the transpose operation that swaps the time and frequency dimension. And, $Mixer$ is defined as follows:

$$Mixer(\mathbf{X}) = \mathbf{X} + \mathbf{W_2}\sigma(\mathbf{W_1}LN(\mathbf{X})) \tag{2}$$

Here σ is the GELU function and LN means Layer Normalization. $\mathbf{W_1}$ and $\mathbf{W_2}$ represent transformation matrix of two dense layers. Furthermore, each MLP block has the same size of the input. This "isotropic" design is most similar to Transformers. We tried pyramidal structure whose deeper blocks have a lower resolution and higher frequency as well, but the result was not too good. In addition, MLP-SVNet does not use any position embeddings because the MLP blocks are sensitive to the order of the input tokens.

3.3. Loss Function

To explicitly enforce the similarity for intra-class samples and the diversity for inter-class samples, several variants based on the Softmax loss function have been proposed, and we have carried out a detailed comparison of the different loss functions in our previous paper [17]. In this paper, we choose the best performance of them, Additive Angular Margin Softmax (AAM-Softmax) [18] will be applied to train the model.

4. EXPERIMENT

4.1. Dataset

The performance of the proposed system MLP-SVNet is assessed by VoxCeleb [19] datasets. VoxCeleb2 development set is used for training. It comprises 1,092,009 utterances among 5,994 speakers, extracted from videos of YouTube. To generate extra training samples and increase the diversity of data, we perform online data augmentation [20] with MUSAN [21] and RIR dataset [22]. The noise type in MUSAN includes ambient noise, music, television, and babble noise for the background additive noise. Augmented data is generated by mixing noise with original speech. For the reverberation, the convolution operation is performed with 40,000 simulated room impulse responses in the RIR dataset. During the training process, we decide whether to augment each sample with the probability 0.6.

We use 40-dimensional filter bank with 25ms windows and 10ms shift as the acoustic features. All MLP-SVNets are trained on chunks of speech features with 300 frames. During the test, we first split each utterance into multiple chunks with 300 frames. Then, we get the embedding for each utterance by averaging the embeddings extracted from these chunks.

4.2. Configuration

During the training, MLP-SVNets are optimized by SAM [23] optimizer with momentum 0.9 and weight decay of $1e$-4. In addition, we adopt AAM-Softmax [18] as loss function for better performance. The scale parameter and the margin of AAM are set to 32 and 0.2 respectively. The whole training process will last 165 epochs while the learning rate decreases from 0.1 to $1e$-5 exponentially. The training is paralleled on 4 GPUs, with the batch-size is set to 64.

4.3. Investigation of Different Patch Methods

Table 1. **Results with different patch methods**. These results are all obtained when the number of MLP blocks is set to 6. Without patch means that not stack the frames with its neighbor contents.

Methods	EER (%)		
	Vox-O	Vox-E	Vox-H
w/o Patch	1.622	1.638	2.707
Patch 1D	**1.361**	**1.469**	**2.492**
Patch 2D	1.819	1.810	3.045

Firstly, we will conduct an investigation of different patching methods in MLP-SVNet, and the results are presented in Table 1. From the table we can find that, patching 1D obtains the best performance on this text-independent speaker verification task among all the results. It shows that stacking the frames with their neighbor contents can better aggregate local information and bring significant performance gain. However, this phenomenon does not appear in the patch 2D method which means that splitting along frequency dimension is not conducive to extracting good speaker information.

As mentioned above, patching 1D outperforms other patch methods with the lowest EER. Based on the patch 1D method, an exploration about the influence of patching size has been conducted in Table 2. According to the results, we find it is necessary to introduce some local information through the patching method, but too large patch size will hurt the performance.

Table 2. **Results with different patch size**. These results are obtained with patch 1D methods and MLP block number is set to 6. Patch size = 1 is equivalent to not stacking the frames with its neighbor contents, which means without patching.

Size	EER (%)		
	Vox-O	Vox-E	Vox-H
1	1.622	1.638	2.707
3	**1.361**	1.469	2.492
5	1.377	**1.447**	**2.444**
7	1.393	1.493	2.518
9	1.435	1.499	2.537

4.4. Investigation of Different MLP Block Number

In our experiments, we also analyzed the effect of MLP block number on the MLP-SVNet. EER results for the different block numbers are presented in Table 3. We see that the increase of the block number only brings a little improvement and MLP-SVNet can achieve a comparable performance even though only a few blocks are used. This benefits from MLP-SVNet's excellent ability of modeling global information. With temporal Mixer blocks, global information is well aggregated and mixed.

Table 3. **Results with different number of MLP blocks**. The results are obtained with patch 1D method and size is set to 3.

Block	EER (%)		
	Vox-O	Vox-E	Vox-H
2	1.659	1.747	2.840
4	1.435	1.496	2.515
6	1.361	**1.469**	**2.492**
8	**1.314**	1.484	2.510

4.5. Comparing with other systems

A performance overview of other speaker verification systems and our proposed MLP-SVNet system is given in Table 4. According to the results, our proposed architecture MLP-SVNet can obviously outperform most of the traditional systems in addition to the state-of-the-art ECAPA-TDNN system which has a more dedicated design to leverage multi-scale information. It reveals that MLP with less inductive bias and more trainable parameters, is superior at capturing long-range dependencies and local features compared to other models based on convolution or self-attention.

Because the proposed MLP-SVNet is totally based on MLP, it has a very different architecture from the models based on convolution or self-attention. We expect that system fusion with the state-of-the-art system can improve performance further. Table 5 presents the results of the different fusion systems. It shows that fusion of ECAPA-TDNN and MLP-SVNet gives the most significant performance gain, which suggests that it can produce the most complementary speaker embeddings comparing with X-vector, R-vector, and S-vector.

Table 4. **Comparison with other speaker verification systems.** We implement all these systems in our experiments.

Systems	EER (%)		
	Vox-O	Vox-E	Vox-H
X-vector [1]	2.117	2.220	3.911
R-vector			
ResNet18 [2]	1.770	1.784	3.020
ResNet34 [2]	1.463	1.555	2.767
S-vector [3]	2.915	2.872	4.754
ECAPA-TDNN [7]	1.080	1.196	2.130
MLP-SVNet	1.361	1.469	2.492

Table 5. **Results of different systems fused with ECAPA-TDNN**. System fusion is based on score weighted summation. ECAPA-TDNN is the state-of-the-art system.

System 1	System 2	EER (%)		
		O	E	H
ECAPA-TDNN	—	1.080	1.196	2.130
	X-vector	1.005	1.173	2.154
	R-vector	1.000	1.190	2.148
	S-vector	1.234	1.340	2.498
	MLP-SVNet	**0.973**	**1.130**	**2.020**

5. CONCLUSION

In this work, we propose a new multi-layer perceptrons based speaker verification network (MLP-SVNet), which doesn't use any convolution or self-attention mechanisms. It applies MLPs across either temporal or frequency for modeling the local and global information at the same time. In the experiments, the results show that MLP-SVNet can significantly outperform X-vector, R-vector, and S-vector. It reveals that MLP-SVNet is superior at capturing long-range dependencies and local features compared to other models. Moreover, benefit from the totally different architecture of MLP-SVNet, it complements the SOTA system well and can further improve the system performance.

6. ACKNOWLEDGEMENTS

This work was supported by the China NSFC projects (No. 62122050 and No. 62071288), and Shanghai Municipal Science and Technology Major Project (No. 2021SHZDZX0102). Experiments have been carried out on the PI super-computer at Shanghai Jiao Tong University.

7. REFERENCES

[1] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. IEEE ICASSP*. IEEE, 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] Pooyan Safari, Miquel India, and Javier Hernando, "Self-attention encoding and pooling for speaker recognition," *Proc. ISCA Interspeech*, pp. 941–945, 2020.

[4] Zhengyang Chen, Shuai Wang, and Yanmin Qian, "Self-supervised learning based domain adaptation for robust speaker verification," in *Proc. IEEE ICASSP*. IEEE, 2021, pp. 5834–5838.

[5] Zhengyang Chen, Shuai Wang, and Yanmin Qian, "Adversarial domain adaptation for speaker verification using partially shared network.," in *Proc. ISCA Interspeech*, 2020, pp. 3017–3021.

[6] Bing Han, Zhengyang Chen, Zhikai Zhou, and Yanmin Qian, "The sjtu system for short-duration speaker verification challenge 2021," *Proc. Interspeech 2021*, pp. 2332–2336, 2021.

[7] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[8] Wei Zhang et al., "Shift-invariant pattern recognition neural network and its optical architecture," in *Proceedings of annual conference of the Japan Society of Applied Physics*, 1988.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[11] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, et al., "Mlp-mixer: An all-mlp architecture for vision," *arXiv preprint arXiv:2105.01601*, 2021.

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

[13] Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip HS Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[14] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141.

[15] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. IEEE ICASSP*, 2014, pp. 4052–4056.

[16] Dan Hendrycks and Kevin Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[17] Xu Xiang, Shuai Wang, Houjun Huang, Yanmin Qian, and Kai Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition," in *Proc. IEEE APSIPA ASC*. IEEE, 2019, pp. 1652–1656.

[18] Yi Liu, Liang He, and Jia Liu, "Large margin softmax loss for speaker verification," *arXiv preprint arXiv:1904.03479*, 2019.

[19] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. ISCA Interspeech*, 2018, pp. 1086–1090.

[20] Weicheng Cai, Jinkun Chen, Jun Zhang, and Ming Li, "On-the-fly data loader and utterance-level aggregation for speaker and language recognition," *IEEE/ACM Trans. ASLP.*, vol. 28, pp. 1038–1051, 2020.

[21] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[22] 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 *Proc. IEEE ICASSP*. IEEE, 2017, pp. 5220–5224.

[23] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," *arXiv preprint arXiv:2010.01412*, 2020.