MFA: TDNN WITH MULTI-SCALE FREQUENCY-CHANNEL ATTENTION FOR TEXT-INDEPENDENT SPEAKER VERIFICATION WITH SHORT UTTERANCES

Tianchi Liu[1,2], *Rohan Kumar Das*[3], *Kong Aik Lee*[1] *and Haizhou Li*[2,4]

[1]Institute for Infocomm Research, A⋆STAR, Singapore
[2]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[3]Fortemedia Singapore, Singapore [4]The Chinese University of Hong Kong, Shenzhen, China

ABSTRACT

The time delay neural network (TDNN) represents one of the state-of-the-art of neural solutions to text-independent speaker verification. However, they require a large number of filters to capture the speaker characteristics at any local frequency region. In addition, the performance of such systems may degrade under short utterance scenarios. To address these issues, we propose a multi-scale frequency-channel attention (MFA), where we characterize speakers at different scales through a novel dual-path design which consists of a convolutional neural network and TDNN. We evaluate the proposed MFA on the VoxCeleb database and observe that the proposed framework with MFA can achieve state-of-the-art performance while reducing parameters and computation complexity. Further, the MFA mechanism is found to be effective for speaker verification with short test utterances.

Index Terms— multi-scale frequency-channel attention, text-independent speaker verification, short utterance

1. INTRODUCTION

Speaker verification (SV) aims to verify the claimed identity of an enrolled speaker, in either a text-dependent or a text-independent manner [1]. The former is based on fixed short phrases [2], while the latter doesn't impose any restrictions on the phonetic content [3]. As the text is fixed in text-dependent SV, only a small amount of data is often used during training [4]. On the contrary, text-independent SV requires a larger amount of training data than text-dependent SV [5], that limits the scope of its applications. Data-efficient text-independent SV for short utterances remains a challenge [4, 6–8].

In the past few decades, SV research witnessed several major breakthroughs. One of the advanced frameworks for SV is x-vector [9, 10], which employs time delay neural networks (TDNNs) to extract frame-level features and finally represent them by fixed-length vectors. Similarly, deep convolutional neural networks (CNN) were exploited and achieved highly competitive performance for SV [11, 12]. Specially, ResNet-based networks have been widely adopted as the backbone for SV system in the recent years [13–19].

Along the similar direction of neural speaker embedding, the ECAPA-TDNN [15] achieves the current state-of-the-art (SOTA) performance by introducing more emphasis on channel attention, propagation and aggregation. However, TDNNs still suffer from the drawback whereby plenty of filters are required to model the speaker characteristics occurring at some local frequency regions [19]. Meanwhile, their performance under short utterance scenario still needs attention.

We propose a multi-scale frequency-channel attention (MFA) module to handle the shortcomings of TDNNs to handle the short utterances for SV in this work. Motivated by ECAPA CNN-TDNN, which further strengthens ECAPA-TDNN by adding a CNN-based front-end to incorporate frequency translational invariance, the MFA module is designed as a front-end module for TDNNs, to obtain high resolution feature representations from short utterances. It employs a frequency-channel attention to assert the ability of effectively capturing speaker information from local regions for TDNNs.

Further, the local and temporal information are modeled by a novel dual-path structure at multiple scales. Meanwhile, by incorporating these two modules, a local cross-channel attention mechanism is developed to contribute towards an efficient framework. The primary contributions of this work are: (1) proposal of frequency-channel attention module, (2) designing dual-pathway for local and temporal information modeling, and (3) local cross-channel attention for the MFA by incorporating (1) and (2), which altogether devote to have data-efficient SV framework for short utterance scenarios.

2. MULTI-SCALE FREQUENCY-CHANNEL ATTENTION

The proposed MFA front-end is shown in Fig. 1, which consists of two parts. The dual-pathway multi-scale (DM) module extracts multi-scale information and the frequency-channel attention (FA) module is used to emphasize important local frequency region. By combining these two modules, we obtain an efficient local cross-channel attention mechanism.

This work is partially supported by Science and Engineering Research Council, A*STAR, Singapore, through the National Robotics Program under Human-Robot Interaction Phase 1 (Grant No. 192 25 00054); in part by A*STAR under its Feasibility Study Scheme (Project No. FS-2021-001); in part by A*STAR under its RIE2020 Advanced Manufacturing and Engineering Domain (AME) Programmatic Grant No. A1687b0033.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

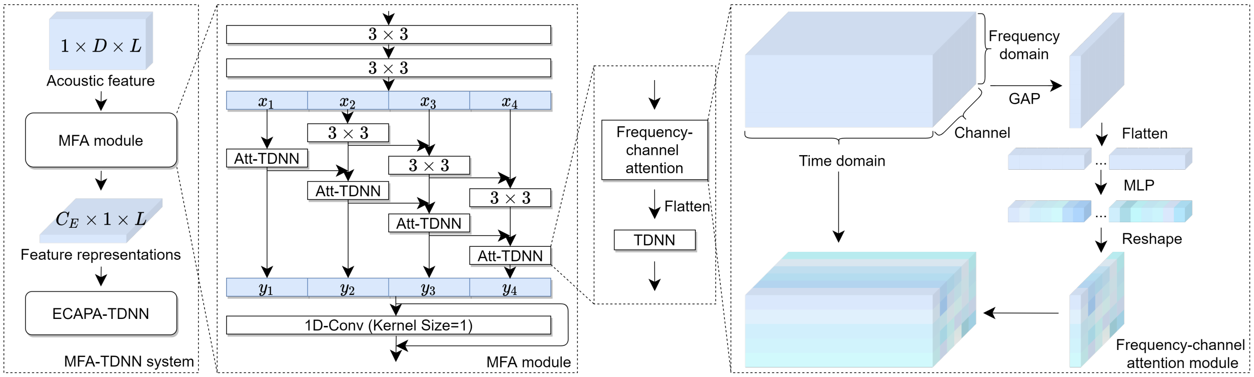

Fig. 1. A network architecture of the proposed MFA-TDNN system with multi-scale frequency-channel attention (MFA) module. The white box with 3 × 3 indicates 2D-CNN with kernel size of 3. The blue boxes indicate intermediate feature maps. We consider the scale dimension s=4 as an example, therefore, the feature maps are equally divided into four groups along the channel dimension. The input acoustic feature is represented as a $1 \times D \times L$ feature map with 1 channel, D dimensional feature frame and L frames. C_E indicates the number of channels in the convolutional frame layers of ECAPA-TDNN.

2.1. Dual-pathway multi-scale module

The Res2Net has the ability to improve the representation by modelling multi-scale information within one layer [20] . Recently, Res2Net is found to be very effective for SV task as well [13,15,19]. In [15], a Res2Net structure is studied where the 2D-CNN is replaced by a TDNN. In this way, the network is expected to process multi-scale features while re-scaling the frame-level features given a wider temporal context over the recording. However, as short utterances offer limited information, it makes sense to extract the speaker characteristics at different scales for an improved speaker characterization.

Motivated by the above, we propose a DM module for dealing with short utterance SV. Considering the fact that the TDNN-based Res2Net module relies on a large number of filters to process well across all the frequency bands and then models speaker characteristics occurring at any local frequency region, we enrich the representation ability of local regions by providing high-resolution details and emphasizing the important frequency bands.

In this work, we divide the feature maps into several groups along the channel dimension C based on the scale dimension factor s. For the first pathway, a 2D-CNN with 3 × 3 filters of n channels ($n = C\,/\,s$) extracts features from group i of input feature x_i. The output features of the previous 2D-CNN are then sent to next group i + 1 of 2D-CNN along with another group of input feature maps x_{i+1}. This pathway is expected to filter the high-resolution features from multi-scale regions. Similarly, the second pathway is designed between a small group of TDNNs to bridging the re-scaled frame-level features with global temporal information. In addition, the frequency channel attention is asserted between 2D-CNN and TDNN for each scale. It is expected to emphasize the specific frequency regions before feeding the feature into TDNN, which will be detailed in the next subsection.

Finally, a 1D-CNN is employed to fuse the information from different scales. The equivalent receptive field increases much more than the classic Res2Net due to the dual-pathway design for any possible path along with, where the input features are processed to output. This will result in many equivalent feature scales due to combination effects [20].

2.2. Frequency-channel attention module

With the advent of deep learning, various attention mechanisms are widely used in SV [15, 19, 21, 22]. Among them, squeeze-and-excitation (SE) block [23] draws great attention due to its success in the field of computer vision, and has been applied to SV as well [15, 19]. The information of some local frequency regions shows great importance for SV. Therefore, we inherit the idea of SE block and propose a novel FA block to emphasize the important information by re-weighting frequency bands and channels.

The channel relationship represents a composition of instance-agnostic functions with local receptive regions [23]. To effectively extract features from the local fields for short utterance SV, we attempt to re-scale local feature responses adaptively by modeling their interdependencies between both channels and frequency bands. Therefore, different from SE block, which only conducts attention on channels, we perform the attention across both channel and frequency dimensions. Specifically, the global temporal information is firstly squeezed into a channel-frequency descriptor as shown in Fig. 1. This is achieved by using global average pooling (GAP). The adaptive re-scaling is performed by employing a multi-layer perceptron (MLP) following the excitation operation of SE block [23]. Finally, the original feature map is re-weighted by the FA map obtained in the previous step.

Prior study shows that an efficient channel attention network (ECA-Net) is achieved by replacing the cross-channel interaction with the local cross-channel attention [24]. By combining the FA module with a DM module in MFA, the channels are divided into multiple scales, thereby limiting the interactions only to adjacent channels. By doing so, we expect to improve the efficiency of the model in a similar way to ECA-Net, by avoiding the interaction across all channels.

3. EXPERIMENTS

3.1. Database

The experiments are conducted on VoxCeleb1 [25] and VoxCeleb2 dataset [26]. Only the development partition of the VoxCeleb2 dataset is used for training, and a small portion of about 2% of it is reserved for validation. On the other hand, the VoxCeleb1 dataset is used for testing. It is noted that there are no overlapping speakers between the development set of VoxCeleb2 and VoxCeleb1 dataset. We evaluate the models on three different popular test set, namely VoxCeleb1-O, VoxCeleb1-E and VoxCeleb1-H. The full-length test utterances are used for when comparing with other SOTA systems.

As the training of neural networks benefits from data augmentation [27], we employ five augmentation techniques to increase the diversity of the training data. The first two follow the idea of random frame dropout in the time domain [27] and speed perturbation [28]. The remaining three are a set of reverberate data, noisy data, and a mixture of the both.

We build a truncated test set of short segments by keeping the original trial pairs, while considering only the first several seconds. We set the short duration cases in the range 4-10s, with a step of 1 s. As the shortest duration of the utterances in VoxCeleb1-O test set is 4s, those testing samples shorter than the set maximum duration are remaining unaltered.

3.2. Systems description

We consider both ECAPA-TDNN [15] and ECAPA CNN-TDNN [19] as baseline systems for the studies. As we re-implement them on our own for the studies, the results of these two systems are marked with '(Re-implemented)' in Section 4. In our systems, we use 80-dimensional filterbank (Fbank) features. We provide details of the two baselines and proposed MFA-TDNN based systems in the following:

- **ECAPA-TDNN (Re-implemented)**: It is the standard ECAPA-TDNN as proposed in [15]. The number of channels is 512 in the convolutional frame layers.
- **ECAPA CNN-TDNN (Re-implemented)**: Four layers of CNN are employed as a front-end for ECAPA-TDNN as proposed in [19]. Different from [19], we do not use a larger version of ECAPA-TDNN, but the standard version mentioned above. This is for fair comparisons with ECAPA-TDNN and proposed MFA-TDNN.
- **MFA-TDNN (Standard)**: The standard ECAPA-TDNN with proposed MFA as a front-end. The scale dimension s is set to 4, while the dimension of channels C for MFA is 32.
- **MFA-TDNN (Lite)**: The lite version of MFA-TDNN, where the number of channels is 480 instead of 512 in the convolutional frame layers of ECAPA-TDNN backend. The scale dimension s is set to 4, whereas the dimension of channels C for MFA is 24.

3.3. Training strategy

The experiments are conducted using SpeechBrain Toolkit[1]. For fair comparisons, all four systems discussed above are trained under the same training strategy following that in [15]. Specifically, the Adam optimizer with cyclical learning rate [31] varying between 1e-8 and 1e-3 following triangular policy [31] is used for training all models. The loss function is additive angular margin softmax (AAM-softmax) [32] with a margin of 0.2 and a scale of 30. In addition, a weight decay used for all the weights in the model is set to 2e-5.

All the samples are cut into 3-seconds segments during training. The mini-batch size is 384 (64 original samples each with 5 augmented samples). A total of 12 epochs are trained for each system. At the end of each epoch, the model is evaluated on the validation set to find the best model for testing.

3.4. Evaluation protocol

We report the performances in terms of equal error rate (EER) and the minimum detection cost function (minDCF) with P_{target} = 0.01 and C_{FA} = C_{Miss} = 1. The test trial scores are produced by calculating the cosine distance between embeddings. The S-norm [33] is applied to normalize the scores.

4. RESULTS AND DISCUSSIONS

Tab. 1 shows the performance comparison of the proposed MFA-TDNN based systems to the two baselines considered in this work. First, we observe that introduction of CNN module in baseline ECAPA-TDNN [15] improves the performance in case of ECAPA CNN-TDNN [19] baseline. On comparing to our proposed MFA-TDNN based systems, we observe that our MFA-TDNN (Standard) performs better than the stronger baseline ECAPA CNN-TDNN in most of the test cases. In addition, the lighter version of MFA-TDNN (Lite) also achieves improved performance than the ECAPA-TDNN baseline. It is noted that MFA-TDNN (Lite) has nearly 1.5M less parameters than MFA-TDNN (Standard). Further, it is worth to mention that both of our MFA-TDNN based proposed system has less parameters and multiply accumulate (MAC) operations than the two baselines. This suggests that our proposed systems not only provide improved results, but they also have compact architecture due to fewer parameters as well as computation complexity.

Let us now compare the proposed systems with other existing SOTA systems. Tab. 1 also shows the performances of several SOTA systems on the VoxCeleb database reported in respective works. We observe that MFA-TDNN outperforms all these SOTA systems. In addition, none of these

[1] https://speechbrain.github.io/

Table 1. Performance in EER(%) and minDCF of the baselines, proposed MFA-TDNN and other SOTA systems on VoxCeleb1 test sets. The MFCC indicates mel frequency cepstral coefficients and Spec. represents spectrogram.

Model	Input Feature	Params	MAC	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
		(Million)	(Giga)	EER	minDCF	EER	minDCF	EER	minDCF
ECAPA-TDNN [15] (Re-implemented)	80-dim Fbank	6.19	1.57	1.0050	0.0991	1.2076	0.1311	2.2593	0.2197
ECAPA CNN-TDNN [19] (Re-implemented)	80-dim Fbank	7.66	2.29	0.9199	0.0921	1.1107	**0.1154**	2.1406	0.2036
Proposed: **MFA-TDNN (Standard)**	80-dim Fbank	**7.32**	**1.91**	**0.8561**	0.0923	**1.0834**	0.1175	**2.0487**	**0.1897**
Proposed: **MFA-TDNN (Lite)**	80-dim Fbank	**5.93**	**1.50**	0.9678	**0.0913**	1.1382	0.1208	2.1740	0.1993
Extended-TDNN (Large) [15]	80-dim MFCC	20.4	-	1.26	0.1399	1.37	0.1487	2.35	0.2153
Extended-TDNN [15]	80-dim MFCC	6.8	-	1.49	0.1604	1.61	0.1712	2.69	0.2419
ResNet18 [15]	80-dim MFCC	13.8	-	1.47	0.1772	1.60	0.1789	2.88	0.2672
ResNet34 [15]	80-dim MFCC	23.9	-	1.19	0.1592	1.33	0.1560	2.46	0.2288
ARET-25 [29]	161-dim Spec.	12.2	-	1.39	-	1.52	-	2.61	-
H/ASP [30]	64-dim Fbank	8.0	-	1.15	-	1.35	-	2.49	-
DDB + Gate [22]	30-dim MFCC	8.83	-	2.31	0.268	-	-	-	-

Table 2. Performance in EER (%) of baselines and proposed MFA-TDNN under short utterance scenario on truncated testset.

System	Duration for test utterances						
	≤ 10 s	≤ 9 s	≤ 8 s	≤ 7 s	≤ 6 s	≤ 5 s	≤ 4 s
ECAPA-TDNN	1.0050	0.9731	0.9784	0.9997	1.0529	1.0901	1.2284
ECAPA CNN-TDNN	0.9199	0.9253	0.9306	0.9465	0.9784	1.0422	1.1858
MFA-TDNN (Standard)	**0.8615**	**0.8508**	**0.8402**	**0.8455**	**0.8615**	**0.9253**	**1.0210**

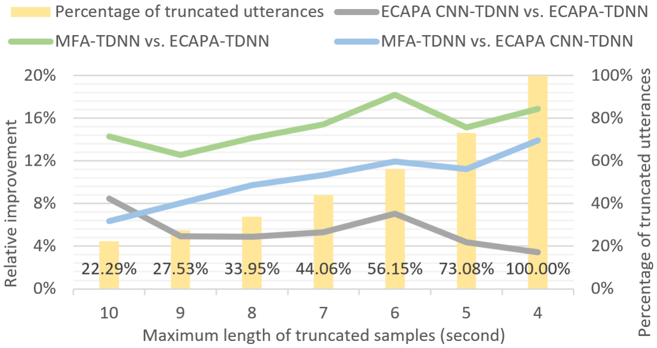

Fig. 2. Relative improvement between different systems for short utterance test sets. The MFA-TDNN (Standard) is used.

SOTA systems is more efficient than MFA-TDNN systems when comparing the number of parameters and MACs. This further shows the advantages of MFA-TDNN systems due to the design of DM and FA modules.

We now compare our proposed MFA-TDNN with the two baseline systems under short utterance test scenario. The experiments are conducted on the truncated data sets as mentioned in Section 3.1 and the results are shown in Tab. 2. All three systems are tested under different settings of maximum duration (4 to 10 seconds). From Tab. 2, we observe that the MFA-TDNN consistently outperforms both the baselines ECAPA-TDNN and ECAPA CNN-TDNN for all the cases of short test utterances considered. It is worth noting that the improvement due to the proposed MFA-TDNN is particularly visible for short utterance scenarios.

In order to better understand the performance trend of these systems under short utterance conditions, we make further analysis in Fig. 2. The curves in the figure represent the relative improvement trend for the three cases as the duration of samples shortens, while the bars indicate the percentage of testing utterances truncated. We observe that under the condition of maximum 10 seconds used for testing, 22.29% testing samples are truncated. On the other hand, all the testing samples need to be shorten for 4-seconds test case. We also observe from Fig. 2 that the relative improvement of ECAPA CNN-TDNN over ECAPA-TDNN (the grey curve) did not improve as the test duration continues to truncate. However, compared to ECAPA-TDNN and ECAPA CNN-TDNN, the relative improvements of the proposed MFA-TDNN is significantly visible as the duration of test utterances reduce. This shows the effectiveness of the proposed MFA-TDNN for text-independent SV with short utterances.

5. CONCLUSION

We proposed a novel multi-scale frequency-channel attention (MFA) framework that efficiently captures the local information and frame-level temporal information by the dual-pathway multi-scale module and emphasize the important frequency region by frequency-channel attention in TDNN systems. Further, by cooperating these two novel modules, local cross-channel attention is achieved to model speaker characteristics from short samples efficiently. Studies on VoxCeleb dataset reveal that we greatly benefit from proposed MFA module. In addition, the further studies on the truncated test sets show the superiority of proposed MFA-TDNN for short sentences compared to various baselines.

6. REFERENCES

[1] Joseph P Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[2] Matthieu Hèbert, "Text-dependent speaker recognition," *Springer-Verlag Heidelberg*, pp. 743–762, 2008.

[3] Tianchi Liu, Rohan Kumar Das, Kong Aik Lee, and Haizhou Li, "Neural acoustic-phonetic approach for speaker verification with phonetic attention mask," *IEEE Signal Processing Letters*, 2022.

[4] Rohan Kumar Das and S. R. Mahadeva Prasanna, "Speaker verification from short utterance perspective: a review," *IETE Technical Review*, vol. 35, no. 6, pp. 599–617, 2018.

[5] Rohan Kumar Das, *Speaker verification using sufficient train and limited test data*, PhD thesis, Septemebr 2017.

[6] Kai Liu and Huan Zhou, "Text-independent speaker verification with adversarial learning on short utterances," in *Proc. IEEE ICASSP*, 2020, pp. 6569–6573.

[7] Gautam Bhattacharya, Md Jahangir Alam, and Patrick Kenny, "Deep speaker embeddings for short-duration speaker verification.," in *Proc. Interspeech*, 2017, pp. 1517–1521.

[8] Zili Huang, Shuai Wang, and Kai Yu, "Angular softmax for short-duration text-independent speaker verification.," in *Proc. Interspeech*, 2018, pp. 3623–3627.

[9] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur, "Deep neural network embeddings for text-independent speaker verification.," in *Proc. Interspeech*, 2017, pp. 999–1003.

[10] Kong Aik Lee, Qiongqiong Wang, and Takafumi Koshinaka, "Xi-vector embedding for speaker recognition," *IEEE Signal Processing Letters*, pp. 1385–1389, 2021.

[11] Tianchi Liu, Maulik C Madhavi, Rohan Kumar Das, and Haizhou Li, "A unified framework for speaker and utterance verification.," in *Proc. Interspeech*, 2019, pp. 4320–4324.

[12] Yi Ma, Kong Aik Lee, Ville Hautamäki, and Haizhou Li, "PL-EESR: Perceptual loss based end-to-end robust speaker representation extraction," in *Proc. IEEE ASRU*, 2021.

[13] Tianyan Zhou, Yong Zhao, and Jian Wu, "ResNeXt and Res2Net structures for speaker verification," in *IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 301–307.

[14] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.

[15] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[16] Dao Zhou, Longbiao Wang, Kong Aik Lee, Yibo Wu, Meng Liu, Jianwu Dang, and Jianguo Wei, "Dynamic margin softmax loss for speaker verification.," in *Proc. Interspeech*, 2020, pp. 3800–3804.

[17] 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," in *Proc. Interspeech*, 2020, pp. 2977–2981.

[18] Ruijie Tao, Rohan Kumar Das, and Haizhou Li, "Audio-visual speaker recognition with a cross-modal discriminative network," in *Proc. Interspeech*, 2020, pp. 2242–2246.

[19] Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck, "Integrating frequency translational invariance in TDNNs and frequency positional information in 2D ResNets to enhance speaker verification," *arXiv preprint arXiv:2104.02370*, 2021.

[20] 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*, pp. 652–662, 2019.

[21] Tianchi Liu, Rohan Kumar Das, Maulik Madhavi, Shengmei Shen, and Haizhou Li, "Speaker-utterance dual attention for speaker and utterance verification," in *Proc. Interspeech*, 2020, pp. 4293–4297.

[22] Yiheng Jiang, Yan Song, Ian McLoughlin, Zhifu Gao, and Li-Rong Dai, "An effective deep embedding learning architecture for speaker verification," in *Proc. Interspeech*, 2019, pp. 4040–4044.

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

[24] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE CVPR*, 2020, pp. 11531–11539.

[25] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.

[26] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[27] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[28] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech*, 2015, pp. 3586–3589.

[29] Ruiteng Zhang, Jianguo Wei, Wenhuan Lu, Longbiao Wang, Meng Liu, Lin Zhang, Jiayu Jin, and Junhai Xu, "ARET: aggregated residual extended time-delay neural networks for speaker verification," in *Proc. Interspeech*, 2020, pp. 946–950.

[30] Yoohwan Kwon, Hee-Soo Heo, Bong-Jin Lee, and Joon Son Chung, "The ins and outs of speaker recognition: lessons from voxsrc 2020," in *Proc. IEEE ICASSP*, 2021, pp. 5809–5813.

[31] Leslie N Smith, "Cyclical learning rates for training neural networks," in *IEEE winter conference on applications of computer vision (WACV)*, 2017, pp. 464–472.

[32] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. IEEE CVPR*, 2019, pp. 4690–4699.

[33] Patrick Kenny, "Bayesian speaker verification with heavy-tailed priors.," in *Proc. Odyssey*, 2010, p. 14.