ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096839

SHORT-SEGMENT SPEAKER VERIFICATION USING ECAPA-TDNN WITH MULTI-RESOLUTION ENCODER

Sangwook Han, Youngdo Ahn, Kyeongmuk Kang, and Jong Won Shin

School of Electrical Engineering and Computer Science
Gwangju Institute of Science and Technology, Gwangju, Korea

ABSTRACT

Time-domain approaches have shown the potential to improve the performance of speaker verification, but still predominant approaches utilize hand-crafted features such as the mel filterbank energies. Although these features are based on speech perception models and exhibited impressive performances, the fixed frame size does not allow good temporal and spectral resolutions at the same time and there is information loss when taking the magnitude spectrum and during frequency rescaling. In this paper, we propose to incorporate multi-resolution time-domain information into the ECAPA-TDNN speaker verification system. We construct a multi-resolution encoder to extract multiple features in different temporal resolutions, and let the extracted features drive the adapter modules. Experimental results showed that the proposed method outperformed other recently proposed approaches when the input length was 2 seconds or shorter for the VoxCeleb dataset. The proposed approach also showed superior performance on the Google Speech Commands dataset v2.

***Index Terms*—** speaker verification, multi-resolution, short-segments, learnable transformation, adapter module

1. INTRODUCTION

Speaker verification (SV) is the process of authenticating the claimed identity of the enrolled speaker given an audio sample. In recent years, a variety of deep neural networks have been proposed as an extractor of speaker embeddings [1–3]. Among them, the ECAPA-TDNN [3] has demonstrated an impressive performance and is adopted in the later researches such as the SV with a self-supervised learning [4] and speech emotion recognition [5]. It utilizes the blocks of Res2Net [6] combined with squeeze and excitation (SE) layers [7], which enhanced the performance by rescaling the feature maps along the channel axis with the weights summarizing information in all frames.

Over the past few decades, perceptually relevant handcrafted features such as mel filterbank energies and mel frequency cepstral coefficients (MFCCs) have been utilized predominantly for SV. These traditional features contain essential information in speech with small dimensions and demonstrated superior performances in SV. However, the features based on the short-time Fourier transform (STFT) with a fixed frame size cannot easily provide good temporal and spectral resolutions at the same time, and the computation of the mel filterbank energies has the loss of information by ignoring the phases and reducing the number of frequency bands. To alleviate these difficulties, several studies [8–12] extracted speaker embeddings directly from the time-domain raw waveforms, and achieved competitive performance [12].

Although many previous approaches were tested with rather long utterances, the application of the SV can be broadened if the system can verify the speaker identity with short utterances. There were several researches that successfully identified speaker using short predefined keywords [1, 2, 13], but the performance of the text-independent speaker recognition with short inputs is significantly worse than that with long utterances [14–17].

In this work, we propose to supplement the ECAPA-TDNN with time-domain multi-resolution features through the adapter modules to extract more informative speaker representation from a short-duration speech segment. We construct the multi-resolution encoder (MRE) which extracts features with various temporal and spectral resolutions by utilizing convolutional kernels of different sizes. The embeddings generated by the encoder with a shorter kernel are fed into the next encoder with a larger kernel to effectively enrich the information in the encoded features for each resolution. The extracted multi-resolution features are used in the adapter module before each SE-Res2Block of the ECAPA-TDNN to transform input features for each block. Experimental results on VoxCeleb dataset showed that our proposed system improved the performance of the short-segment SV with the input length of 2 second or less significantly. It is also shown that the proposed system exhibited superior performance on the Google Speech Commands dataset v2 consisting of 1-second-long utterances in a text-independent configurations.

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2021-0-01835) supervised by the IITP (Institute of Information Communications Technology Planning Evaluation) and Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01842, Artificial Intelligence Graduate School Program (GIST)).

Authorized licensed use limited to: Kwangju Institute of Science and Technology. Downloaded on May 19,2023 at 16:25:29 UTC from IEEE Xplore. Restrictions apply.

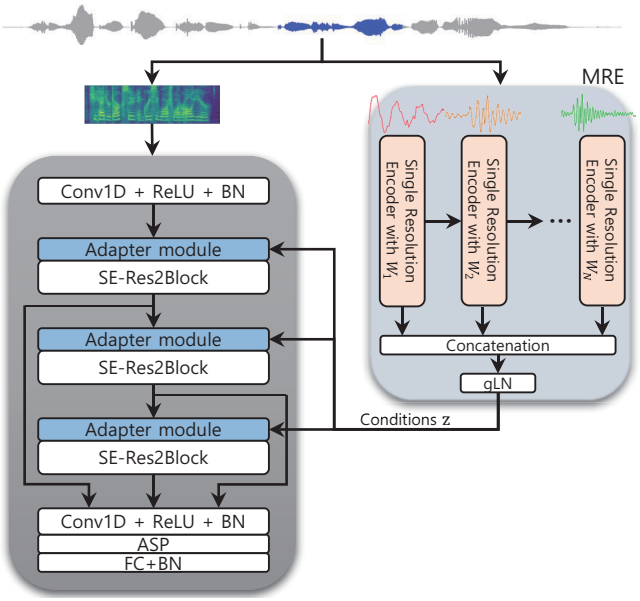

Fig. 1. *Overall block diagram of the proposed ECAPA-TDNN speaker verification system with a multi-resolution encoder.*

2. PROPOSED ECAPA-TDNN WITH MULTI-RESOLUTION ENCODER

The overall block diagram of the proposed system is shown in Fig. 1, which consists of the ECAPA-TDNN with adapter modules and the MRE that extract the multi-resolution features used in the adapter modules. The whole system is jointly trained to minimize a single SV loss in an end-to-end manner. Detailed descriptions of the newly added modules are given in the following subsections.

2.1. Multi-resolution encoder

For the time-frequency analysis using the STFT with a fixed overlap ratio, it is well-known that the longer window provides good spectral resolution with poor temporal resolution and vice versa. To extract information from a short segment as much as possible, we employ multiple encoders with different convolutional kernel sizes and feed embeddings from the encoder with a shorter kernel to the one with a longer kernel, as shown in Fig. 2.

The architecture of each of the single resolution encoder (SRE) is similar to the feature extraction module in [10]. In the n-th SRE, a raw waveform input of length L, $\mathbf{x} \in \mathbb{R}^{1\times L}$, is convolved with H kernels of length W_n and stride $W_n/2$ to produce $\mathbf{x}_n^{'} \in \mathbb{R}^{H\times \frac{2L}{W_n}}$, and then rectified linear unit (ReLU) and layer normalization (LN) [18] are applied. The kernel size W_n is doubled in the next SRE, i.e., $W_n = 2^{n-1}W_1$. The number of channels is changed from H to P using 1×1 convolution to make $\mathbf{x}_n^{''} \in \mathbb{R}^{P\times \frac{2L}{W_n}}$. $\mathbf{x}_n^{''}$ is then added to the embedding from the preceding SRE with W_{n-1} and fed into the temporal convolutional network (TCN) block consisting of the 1-D ConvSE blocks as in [10]. Inside the TCN block, there are three 1-D ConvSE blocks with exponentially increasing dilation factors dependent on n, i.e., $d_i = 2^{n-1}\times 2^{i-1}$, for the i-th block. The dilation factors are configured to cover a large receptive field with a limited number of parameters. The output of TCN block, $\mathbf{z}_n^{'} \in \mathbb{R}^{P\times \frac{2L}{W_n}}$ is then convolved with Q kernels of size M_n and stride $M_n/2$, and ReLU and LN are applied to produce the output of the n-th SRE, $\mathbf{z}_n \in \mathbb{R}^{Q\times \frac{4L}{W_n M_n}}$. We configured M_n to satisfy $\frac{W_n}{2}\cdot\frac{M_n}{2} = S$ in which S is the frame shift to compute the mel filterbank energies, which makes all the $\mathbf{z}_n$ have the same temporal dimension with the mel filterbank energy features. $\mathbf{z}_n^{'}$ is also added to $\mathbf{x}_{n+1}^{''} \in \mathbb{R}^{P\times \frac{2L}{W_{n+1}}}$ in the $(n+1)$-th SRE after matching the dimension by the max pooling operation with a downsampling factor of 2.

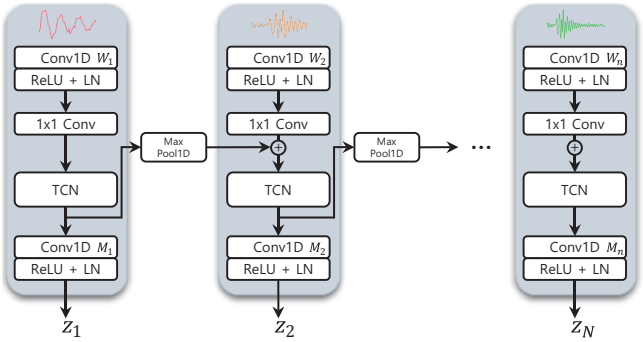

Fig. 2. *Architecture of the multi-resolution encoder (MRE).*

The outputs of the N SREs are then concatenated along the channel axis and then global layer normalization (gLN) [19] is applied to produce the output of the MRE, i.e.,

$$\mathbf{z} = \mathbf{gLN}(\mathbf{Concat}(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N)) \in \mathbb{R}^{NQ\times \frac{L}{S}}, \quad (1)$$

as shown in Fig. 1. $\mathbf{z}$ is then used in the adapter modules before the SE-Res2Blocks as conditions.

2.2. Adapter module

One of the most straightforward ways of conditioning is to sum up the conditioning variable and the intermediate feature in the backbone system, after matching the dimension. The channel dimension of $\mathbf{z}$ can be changed to the number of channels in the input of the SE-Res2Block $\mathbf{h} \in \mathbb{R}^{C\times \frac{L}{S}}$ by applying a 1×1 convolution and then $\mathbf{z}$ can be added to $\mathbf{h}$.

A more sophisticated way may be to extract global and local features from the conditioning variable $\mathbf{z}$ and apply affine transforms of the intermediate feature $\mathbf{h}$ as illustrated in Fig. 3. Fig. 3 shows the structure of the adapter module we place before each SE-Res2Block as shown in Fig. 1, which is similar to the multi-scale channel attention module in [20]. Given a multi-resolution feature $\mathbf{z} \in \mathbb{R}^{NQ\times \frac{L}{S}}$, the global and local

Authorized licensed use limited to: Kwangju Institute of Science and Technology. Downloaded on May 19,2023 at 16:25:29 UTC from IEEE Xplore. Restrictions apply.

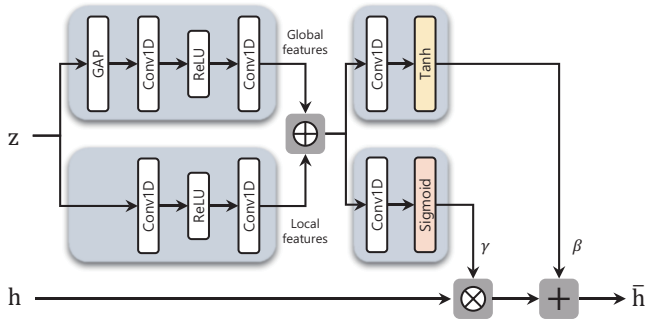

Fig. 3. *The structure of the adapter module.*

features are extracted using bottleneck structures. To extract the global features $\mathbf{z}_g \in \mathbb{R}^{NQ\times 1}$, the global average pooling (GAP) is applied along the temporal axis and then the number of channels is reduced to NQ/r in the first convolutional layer with a ReLU activation and then restored to NQ by the second convolutional layer, where r is the reduction ratio of the bottleneck structure. The local features $\mathbf{z}_l \in \mathbb{R}^{NQ\times \frac{L}{S}}$ are generated in a similar manner except that GAP was not applied. The extracted global and local features are integrated by the broadcasting addition $\oplus$, i.e., $\mathbf{z}_i = \mathbf{z}_g \cdot \mathbf{1}^T + \mathbf{z}_l \in \mathbb{R}^{NQ\times \frac{L}{S}}$ in which $\mathbf{1} \in \mathbb{R}^{\frac{L}{S}\times 1}$ is the all one vector. $\mathbf{z}_i$ is then converted to $\boldsymbol{\gamma} \in \mathbb{R}^{C\times \frac{L}{S}}$ and $\boldsymbol{\beta} \in \mathbb{R}^{C\times \frac{L}{S}}$ using 1×1 convolutions with sigmoid and hyperbolic tangent activations, respectively, which are used to apply element-wise affine transformations to $\mathbf{h}$:

$$\bar{\mathbf{h}} = Adapter(\mathbf{h}, \mathbf{z}) = \boldsymbol{\gamma} \otimes \mathbf{h} + \boldsymbol{\beta} \tag{2}$$

where $\otimes$ denotes the element-wise multiplication. $\bar{\mathbf{h}}$ is then fed into the subsequent SE-Res2Block as input.

3. EXPERIMENTAL SETUP

3.1. Datasets

We used the development set of the VoxCeleb2 [21] for training, which consisted of over 1 million utterances from 5,994 speakers. To evaluate the performance, we used the VoxCeleb1-O, VoxCeleb1-E, and VoxCeleb1-H test sets [21], in which the numbers of speakers are 40, 1,251, and 1,190, respectively. In addition, to evaluate the performance on very short words, we used the Google Speech Commands dataset v2 [22]. The dataset consists of 105,829 1-second-long utterances of 35 words spoken by 2,618 speakers. To perform SV tasks, the dataset was divided into the training, validation, and test sets of 2,356, 27, and 235 speakers, respectively, without overlapping speakers. To validate and evaluate systems, 10,000 and 100,000 pairs were constructed for validation and test sets by choosing a random utterance and finding another utterance from the same or different speaker for each positive or negative pair. The sampling rate was 16kHz for all data.

Table 1. *Speaker verification performances depending on the input length for the VoxCeleb1-O.* † *: re-implementation.*

Systems	Input feature	EER (%)			
		full	5s	2s	1s
MSEA-FPM [15]	40-dim Fbank	1.98	2.17	3.38	5.92
ResNet34-GAP [17]	80-dim Fbank	1.95	-	3.13	6.38
ResNet34-ANF [23]	40-dim Fbank	1.91	2.04	2.88	4.49
ECAPA-TDNN† [3]	80-dim Fbank	1.03	1.05	1.76	3.04
RawNet2 [8]	Waveform	2.43	2.64	3.88	7.24
RawNeXt [11]	Waveform	1.29	1.45	2.34	4.37
Proposed ECAPA-TDNN + MRE	80-dim Fbank + Waveform	**1.01**	**1.03**	**1.32**	**2.33**

3.2. Implementation details

We used 2-second chunks that randomly cropped from each utterance as input, and the 80-dimensional log mel-filterbank with a 25ms window and 12.5ms frame shift was used to compute mel filterbank energies. The data augmentation was applied using the simulated room impulse response (RIR), MUSAN noises [24], and speed perturbation with a factor of 0.9 and 1.1. Voice activity detection (VAD) was not applied. The objective function was AAM-Softmax [25] with a margin of 0.2 and a scale of 30. The Adam optimizer was applied and the learning rate was initialized to $1e^{-3}$ and decayed by a rate of 0.97 every epoch. We trained the model for 100 epochs and then used the resultant model for evaluation. As for the parameters for the MRE, W_1 and M_1 were set to 50 and 16 to match the frame shift of $S = 200$ samples, and the numbers of kernels H, P and Q were set to $\{256, 128, 64\}$. For both the baseline ECAPA-TDNN system and the proposed method, C was set to 1,024. The kernel sizes of all convolutions in adapter modules were 3. The performances were measured in terms of the equal error rate (EER) and the minimum of the detection cost function (MinDCF) with $P_{target} = 0.05$ and $C_{FalseAlarm} = C_{Miss} = 1$.

4. RESULTS AND ANALYSIS

Table 1 shows the comparison of the EERs with the systems recently proposed for short-segment SV for the VoxCeleb1-O dataset. To evaluate the performance on short segments, we used full-length enrollment utterances and short test utterances cropped from the middle of utterances. We can see that the re-implemented ECAPA-TDNN outperformed other spectral- and time-domain approaches which reported performances for short segments for all input lengths. The proposed approach adopting the MRE improved the performance of the ECAPA-TDNN especially for the inputs of 2 seconds and one second by 25% and 23% relatively, respectively.

In addition, we carried out an ablation study on the contribution of the MRE and the adapter module to the performance improvement. As for the effect of the MRE, we have tested temporal encoders with 1 to 4 SREs, while the baseline ECAPA-TDNN correspond to 0 SRE. As an alternative to the

Authorized licensed use limited to: Kwangju Institute of Science and Technology. Downloaded on May 19,2023 at 16:25:29 UTC from IEEE Xplore. Restrictions apply.

Table 2. *Speaker verification performances for the baseline ECAPA-TDNN and the proposed system depending on the configurations of the multi-resolution encoder and the conditioning method for VoxCeleb1-O, VoxCeleb-E and VoxCeleb-H.*

Systems	Convolutional kernel sizes	VoxCeleb1-O						VoxCeleb1-E						VoxCeleb1-H					
		EER(%)			MinDCF			EER (%)			MinDCF			EER(%)			MinDCF		
		2s	1.5s	1s	2s	1.5s	1s	2s	1.5s	1s	2s	1.5s	1s	2s	1.5s	1s	2s	1.5s	1s
Baseline	-	1.76	2.12	3.04	0.1387	0.1544	0.2184	1.86	2.13	2.95	0.1290	0.1411	0.1841	3.45	3.84	5.01	0.1998	0.2205	0.2993
Proposed with summation conditioning	50	1.69	1.91	2.76	0.1326	0.1417	0.1989	1.83	2.08	2.78	0.1218	0.1380	0.1764	3.38	3.79	4.84	0.2014	0.2270	0.2824
	50, 100	1.65	1.88	2.70	0.1297	0.1409	0.1928	1.80	2.04	2.69	0.1208	0.1352	0.1737	3.26	3.67	4.70	0.1961	0.2202	0.2789
	50, 100, 200	1.56	1.86	2.61	0.1141	0.1313	0.1893	1.71	1.98	2.61	0.1134	0.1302	0.1668	3.02	3.61	4.62	0.1924	02193	0.2757
	50, 100, 200, 400	1.43	1.72	2.41	0.1083	0.1178	0.1734	1.64	1.87	2.48	0.1065	0.1208	0.1571	2.92	3.38	4.35	0.1813	0.2074	0.2640
Proposed with adapter module	50	1.69	1.96	2.79	0.1201	0.1463	0.1935	1.76	2.02	2.70	0.1158	0.1350	0.1759	3.23	3.64	4.73	0.1947	0.2202	0.2812
	50, 100	1.58	1.86	2.69	0.1192	0.1283	0.1796	1.65	1.88	2.47	0.1083	0.1261	0.1575	2.99	3.40	4.33	0.1848	0.2086	0.2655
	50, 100, 200	1.53	1.80	2.57	0.1148	0.1271	0.1749	1.62	1.83	2.44	0.1072	0.1211	0.1599	2.96	3.36	4.29	0.1831	0.2052	0.2611
	50, 100, 200, 400	**1.32**	**1.60**	**2.33**	**0.1036**	**0.1260**	**0.1723**	**1.52**	**1.74**	**2.35**	**0.1012**	**0.1158**	**0.1513**	**2.83**	**3.22**	**4.19**	**0.1723**	**0.1947**	**0.2524**

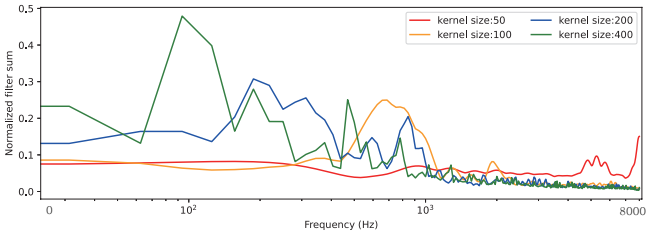

Fig. 4. *The cumulative frequency responses (CFRs) of learned filters with different temporal resolutions in the MRE. The CFRs are normalized by 2-norm for better visualization.*

adapter module, conditioning through simple summation described in the first paragraph in subsection 2.2 was tested. The EER and MinDCF for the VoxCeleb1-O, VoxCeleb1-E, and VoxCeleb1-H datasets when the input length was 2s, 1.5s and 1s are present in Table 2. In all cases, employing an additional SRE improved the performances. It can also be observed that the conditioning with the adapter module enhanced the performance further, while requiring more parameters and computation compared with simple summation. From the results, we can verify that the multiple time-domain encoders with different temporal resolutions were advantageous to extract supplementary information from short segments and the information from the MRE was effectively incorporated to the ECAPA-TDNN through the adapter module.

To further analyze the role of each SRE, we have evaluated the cumulative frequency responses, F_{cum}, of the H learned filters with the kernel size W_n in the first convolutional layer of each SRE defined as

$$F_{cum} = \frac{1}{H}\sum_{i=1}^{H}\frac{F_i}{||F_i||_2} \tag{3}$$

where F_i is the magnitude response of the i-th filter. As expected, the filters with large kernel sizes focused on the lower frequencies, while the shortest filter covered the remaining information in all frequencies.

Finally, to validate the proposed method for another database with short utterances, we have tested it and the baseline ECAPA-TDNN for the Google Speech Commands (GSC) dataset v2. We have examined two different training strategies; the former was to train the model using the GSC dataset with random initialization, and the latter was to pre-train the model using the VoxCeleb2 dataset and fine-tune it with the GSC dataset. For both of the cases, the models were trained with the GSC datasets for 20 epochs with the initial learning rate of $2e^{-4}$, which were enough for models to converge. The results are shown in Table 3. We can see that pre-training with VoxCeleb2 helped to improve the performance, and the performance improvement of the proposed method over the baseline ECAPA-TDNN was even bigger when the training or fine-tuning dataset was composed of 1-second-long utterances compared with the previous experiment in which models were trained with 2-second-long excerps. It may be because 1 second of input was too short to extract enough speaker information from mel filterbank energies only and the role of the MRE was more crucial.

Table 3. *Performances on the Google Speech Commands dataset v2.*

Systems	Training database	EER (%)	MinDCF
Baseline	Speech Commands	8.39	0.4966
	VoxCeleb2 / Speech Commands	6.52	0.3838
Proposed	Speech Commands	2.17	0.1383
	VoxCeleb2 / Speech Commands	**1.76**	**0.1175**

5. CONCLUSION

We propose to incorporate the time-domain multi-resolution encoder into the ECAPA-TDNN for short-segment SV. The MRE extracts features by using multiple SREs with different convolutional kernel sizes and by feeding information from the SRE with a shorter kernel to the one with a longer kernel. The extracted multi-resolution features are used as conditions in the adapter modules to modify the input to each block in the ECAPA-TDNN. Experiments on the VoxCeleb dataset and the Google Speech Commands dataset showed that the proposed system improved the performance of the ECAPA-TDNN for the inputs less than 2 seconds significantly.

Authorized licensed use limited to: Kwangju Institute of Science and Technology. Downloaded on May 19,2023 at 16:25:29 UTC from IEEE Xplore. Restrictions apply.

6. REFERENCES

[1] E. Variani, X. Lei, E. McDermott, IL. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. ICASSP*. IEEE, 2014, pp. 4052–4056.

[2] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proc. ICASSP*. IEEE, 2016, pp. 5115–5119.

[3] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[4] 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," in *Proc. ICASSP*. IEEE, 2022, pp. 6147–6151.

[5] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech emotion recognition using self-supervised features," in *Proc. ICASSP*. IEEE, 2022, pp. 6922–6926.

[6] S. Gao, M.-M. Cheng, M.-H. Yang K. Zhao, X. Zhang and, and P. H. S. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE TPAMI*, vol. 43, no. 2, pp. 652–662, 2019.

[7] J. Hu, L Shen, and G.Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF CVPR*, 2018, pp. 7132–7141.

[8] J. Jung, S. Kim, H. Shim, J. Kim, and H.-J Yu, "Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms," in *Proc. Interspeech*, 2020, pp. 1496–1500.

[9] G. Zhu, F. Jiang, and Z. Duan, "Y-vector: Multiscale waveform encoder for speaker embedding," in *Proc. Interspeech*, 2021, pp. 96–100.

[10] S. Han, J. Byun, and J. W. Shin, "Time-domain speaker verification using temporal convolutional networks," in *Proc. ICASSP*. IEEE, 2021, pp. 6688–6692.

[11] J. Kim, H. Shim, J. Heo, and H.-J Yu, "Rawnext: Speaker verification system for variable-duration utterances with deep layer aggregation and extended dynamic scaling policies," in *Proc. ICASSP*. IEEE, 2022, pp. 7647–7651.

[12] J. w. Jung, Y. Kim, H.-S. Heo, B.-J. Lee, Y. Kwon, and J. S. Chung, "Pushing the limits of raw waveform speaker recognition," in *Proc. Interspeech*, 2022, pp. 2228–2232.

[13] Zhang P, P. Hu, and Xueliang Zhang, "Deep embedding learning for text-dependent speaker verification.," in *Proc. Interspeech*, 2020, pp. 3461–3465.

[14] A. Hajavi and A. Etemad, "A deep neural network for short-segment speaker recognition," in *Proc. Interspeech*, 2019, pp. 2878–2882.

[15] Y. Jung, S. M. Kye, Y. Choi, M. Jung, and H. Kim, "Improving multi-scale aggregation using feature pyramid module for robust speaker verification of variable-duration utterances," in *Proc. Interspeech*, 2020, pp. 1501–1505.

[16] Y. Jung, Y. Choi, H. Lim, and H. Kim, "A unified deep learning framework for short-duration speaker verification in adverse environments," *IEEE Access*, vol. 8, pp. 175448–175466, 2020.

[17] J.-H. Choi, J.-Y. Yang, and J.-H. Chang, "Short-utterance embedding enhancement method based on time series forecasting technique for text-independent speaker verification," in *ASRU*. IEEE, 2021, pp. 130–137.

[18] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[19] Y. Luo and N Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM TASLP*, vol. 27, no. 8, pp. 1256–1266, 2019.

[20] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proc. IEEE/CVF CVPR*, 2021, pp. 3560–3569.

[21] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[22] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[23] S. M. Kye, J. S. Chung, and H. Kim, "Supervised attention for speaker recognition," in *SLT*. IEEE, 2021, pp. 286–293.

[24] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[25] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF CVPR*, 2019, pp. 4690–4699.

Authorized licensed use limited to: Kwangju Institute of Science and Technology. Downloaded on May 19,2023 at 16:25:29 UTC from IEEE Xplore. Restrictions apply.