RAWNEXT: SPEAKER VERIFICATION SYSTEM FOR VARIABLE-DURATION UTTERANCES WITH DEEP LAYER AGGREGATION AND EXTENDED DYNAMIC SCALING POLICIES

Ju-ho Kim, Hye-jin Shim, Jungwoo Heo, and Ha-Jin Yu[*]

School of Computer Science, University of Seoul

ABSTRACT

Despite achieving satisfactory performance in speaker verification using deep neural networks, variable-duration utterances remain a challenge that threatens the robustness of systems. To deal with this issue, we propose a speaker verification system called *RawNeXt* that can handle input raw waveforms of arbitrary length by employing the following two components: (1) A deep layer aggregation strategy enhances speaker information by iteratively and hierarchically aggregating features of various time scales and spectral channels output from blocks. (2) An extended dynamic scaling policy flexibly processes features according to the length of the utterance by selectively merging the activations of different resolution branches in each block. Owing to these two components, our proposed model can extract speaker embeddings rich in time-spectral information and operate dynamically on length variations. Experimental results on the VoxCeleb1 test set consisting of various duration utterances demonstrate that RawNeXt achieves state-of-the-art performance compared to the recently proposed systems. Our code and trained model weights are available at `https://github.com/wngh1187/RawNeXt`.

***Index Terms*—** speaker verification, deep layer aggregation, dynamic scaling policy, short duration, raw waveform

1. INTRODUCTION

Speaker verification (SV) is the task of determining whether the identity of an anonymous voice matches the target speaker. In general, SV is performed as a series of processes: extracting fixed-dimensional utterance-level features from utterances and calculating the similarities between the features. Herein, the utterance-level features (*i.e.*, speaker embeddings) are usually extracted through the network trained by the embedding learning methods. Due to advances in deep learning, deep neural network (DNN)-based embedding learning approaches such as d-vector [1] and x-vector [2] outperform traditional schemes (*e.g.*, i-vector [3]). Although these methods have greater potential, they exhibit unsatisfactory performance for short input utterances [4].

In the field of SV, short utterances are one of the well-known performance degradation factors, increasing the uncertainty of embedding owing to insufficient speaker-specific information [5]. To tackle this challenge, several studies have focused on how to effectively use the sparse speaker information contained in short utterances [6, 7]. They extracted speaker embeddings in a multi-scale aggregation (MSA) manner that adequately fuses and utilizes intermediate features of various time scales within the network. The MSA approach and its variants have shown superior performance for variable-duration utterance SV tasks [6–8].

Inspired by the MSA to short utterances, we aim to advance the embedding extraction process by aggregating features in a more iterative and hierarchical fashion. To achieve this goal, we propose applying deep layer aggregation (DLA) [9] as a speaker embedding extractor. DLA consists of two structures: iterative deep aggregation (IDA) and hierarchical deep aggregation (HDA). Like the MSA, IDA enriches the temporal information by merging features of different time scales from the previous stage (red lines in Fig. 1 (a)). HDA fuses the channel axis of features from different blocks (yellow boxes in Fig. 1 (a)). Typically, channels in a feature map ($\in \mathbb{R}^{T\times C}$) yielded from the 1d convolutional layer contain spectral information [10]. Thus, HDA enhances the spectral information to extract more informative embeddings [11]. Consequently, speaker embedding obtained by aggregating temporal and spectral information using DLA is expected to further improve the SV performance for variable-duration utterances.

Meanwhile, the majority of SV systems process utterances in a fixed way with a series of manually designed layers [5–8]. However, it cannot be guaranteed to be optimal for variable-duration utterances. Therefore, the SV systems require dynamic speaker embedding extraction that can flexibly handle utterances of various lengths. Elastic [12] was proposed for the scale variation of images, and it alleviated this issue by adding downsampling paths to the blocks. Elastic let the network dynamically process data by utilizing the appropriate resolution branches (original or downsampling paths) in each block. In this study, to process features according to the length of the utterance, we introduce an extended dynamic scaling policy (EDSP) based on the Elastic. Compared with the existing method, EDSP increases the resolution branch (upsampling paths) to provide more varied scaling options. In addition, the proposed method exploits a multi-head attention-based gate module to selectively aggregate the activation of each path. Thus, the EDSP encourages the model to better extract speaker information by dynamically responding to the length of utterances.

In summary, DLA enriches the speaker's time-spectral information and EDSP induces flexible operation according to the length of input utterances. By applying both methods, we finally propose an SV system called *RawNeXt* (suggesting the *next* version of [13]), which is robust to length variation of utterances with raw waveforms as inputs. To train and evaluate the models, the VoxCeleb datasets [14, 15] were used. As a result of the experiments, RawNeXt reported state-of-the-art performance for short-utterance SV tasks. Moreover, our proposed system showed results comparable to the top-three single systems of the 2020 VoxCeleb Speaker Recognition Challenge (VoxSRC-20) [16]. Additionally, we demonstrated the effectiveness of RawNeXt's components through ablation and analysis experiments.

2. BASELINE

In SV research, novel approaches are emerging that deal directly with raw waveforms [10, 13] rather than engineered acoustic features such as mel-filterbank energies [1, 2]. It is known that systems trained in a data-driven manner on less-processed data can extract discriminative representations suitable for SV tasks with minimal hyper-parameter search of acoustic feature pre-processing [10, 13]. To take advantage of these pros, we use raw waveforms as model inputs.

[*] Corresponding author.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT Future Planning (2020R1A2C1007081)

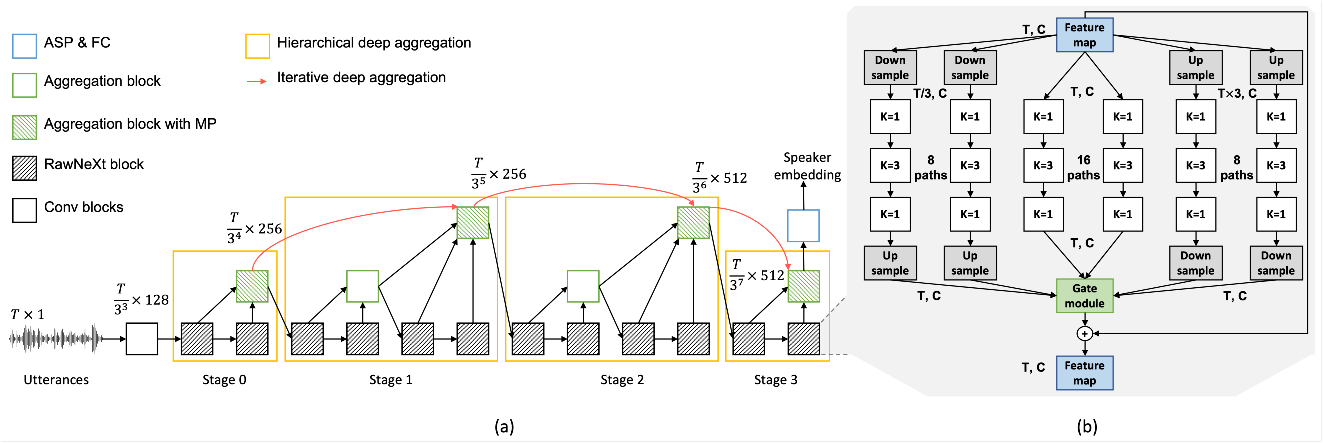

Fig. 1: (a): Overall architecture of RawNeXt. RawNeXt is trained to extract speaker embeddings rich in temporal and spectral information from utterances. Iterative aggregation merges previous shallow stage to progressively propagate features of different resolutions. Hierarchical aggregation combines different channels of blocks in stages to better refine the features. (b): Structure of RawNeXt block. The RawNeXt block divides the original paths in half and additionally processes the input into low- and high-resolution paths. The features calculated for branches of each resolution are restored to their original resolution and aggregated by the gate module. Through these processes, blocks can learn a dynamic scaling policy according to the input data.

Table 1: The architecture of the baseline. The sample size of the input waveform is 59,049. For convolutional layers, numbers inside parentheses refer to the kernel length, stride size, and number of kernels.

Level	Block structure	# Blocks	Output
Convs	Conv(3, 3, 128)	1	2,187×128

	Conv(3, 1, 128)
	Maxpool(3)
		2	

Stage 0	Conv(1,1,256)
	Conv(3,1,256), C=32
	Conv(1,1,256)
	Maxpool(3)
		2	729×256

Stage 1		4	243×256

Stage 2	Conv(1,1,512)
	Conv(3,1,512), C=32
	Conv(1,1,512)
	Maxpool(3)
		4	81×512

Stage 3		2	27×512
Pooling	ASP	1	1,024
Embedding	FC(512)	1	512

Table 1 shows the baseline structure of this study based on ResNeXt [17]. The ResNeXt contains the grouped convolutional layers known as a split-transform-merge strategy in blocks and recently reported reliable performance in SV [18]. Cardinality (C) refers to the number of groups in grouped convolution operations. Stages 0, 1 and 2, 3 have identical block structures, respectively, and max pooling (MP) is applied at the end of each stage. Batch normalization (BN) and ReLU activation are employed after every convolutional layer. Finally, speaker embeddings with 512 dimensions are extracted through the attentive statistical pooling (ASP) [19] and fully connected (FC) layers.

3. PROPOSED METHODS

3.1. Deep layer aggregation

Beyond constructing deeper and wider DNNs to increase accuracy, it is also being explored to connect blocks more closely [22]. By merging features of several layers, systems can yield context-rich representations for target tasks and mitigate the gradient vanishing problem by back-propagating earlier to lower layers [22,23]. Furthermore, in the field of SV, it is known that the embeddings extracted from MSA-based models are robust to short-duration utterances [6, 8]. Similar to the direction of previous studies, we intend to derive speaker embeddings by fusing features in a more iterative and hierarchical manner for utterances of various lengths. To achieve this aim, we use the DLA [9] as a speaker embedding extractor.

Fig. 1 (a) illustrates the overall architecture of the proposed system, RawNeXt, and the multiplicative numbers above each block indicate the size (time × channel) of the corresponding feature. Compared with the baseline as in Table 1, RawNeXt additionally utilizes IDA and HDA modules for feature aggregation at each stage. IDA iteratively merges stages, from shallow to deep. In this way, aggregations of different time resolutions enrich temporal context information in deep features. HDA hierarchically fuses blocks in a tree-structured fashion for each stage. Hence, the context information in spectral domain is enhanced by combining the feature channels of different levels.

The aggregation blocks (green boxes in Fig. 1 (a)) learn to select important information from the multiple inputs and project it into a single output. The aggregation block, N, is formulated as follows:

$$N(x_1,...,x_n) = \sigma(Conv([x_1,...,x_n]))) \qquad (1)$$

where x_i is the output of the previous i-th block, and these outputs are concatenated, denoted as $[\cdot]$. After that, it is transformed into the single output via a convolution layer with a kernel and stride size of 1, followed by BN and ReLU activation functions denoted as $\sigma(\cdot)$. Finally, the last feature output from Stage 3, containing the compressed network-wide information, is converted into speaker embedding through ASP and FC.

3.2. Extended dynamic scaling policy

The representations that do not consider the scale variation of the data are often sub-optimal for the target tasks [24]. To mitigate this issue, Wang *et al.* [12] proposed a method, Elastic. This approach allows the network to learn a scaling policy from data, in which the system can decide among the original and the downsampling paths in each block. Thus, Elastic encourages the network to perform dynamically on the scale of data. Inspired by this scheme, we argue that the variable-duration SV task should also perform flexibly depending on the utterance lengths. Therefore, we propose an EDSP based on Elastic for arbitrary length utterances.

Table 2: Results of comparison with recently proposed speaker verification system for short utterances. ($^{\top}$: drawn from [20], † : our implementation, *: data augmentation)

Model	Input Feature	Loss Function	Aggregation Method	Vox1-O			
				EER% 1s	EER% 2s	EER% 5s	EER% / C_{det}^{min} full
MSEA+FPM [8]	MFB-64	A-Softmax	LDE	5.92	3.38	2.17	1.98 / 0.205
ResNet34 [21]$^{\top}$	MFB-40	Softmax+PN	TAP	4.77	3	2.2	2.08 / 0.234
ResNet34 [20]	MFB-40	Softmax+PN	ANF	4.49	2.88	2.04	1.91 / 0.221
RawNet2 [13]†	Waveform	Softmax	ASP	7.24	3.88	2.64	2.43 / 0.236
ResNeXt (Baseline)	Waveform	Softmax	ASP	6.12	3.68	2.45	2.16 / 0.187
RawNeXt (Proposed)	Waveform	Softmax	ASP	**4.47**	**2.58**	**1.72**	**1.54 / 0.166**
RawNeXt*	Waveform	AAM-Softmax	ASP	**4.37**	**2.34**	**1.45**	**1.29 / 0.142**

Fig. 1 (b) shows the structure of RawNeXt block, in which the EDSP strategy is applied to the baseline block. The proposed EDSP reduces the original-resolution branches from 32 to 16 and extends the feature resolution range by adding eight downsampling and eight upsampling branches in parallel. Features are processed through convolutional layers of the same structure in each branch. Herein, applying the convolution with the same kernel and stride size in different resolutions implies extracting features with receptive fields of different sizes. That is, for the same input, the receive field sizes of the original path, downsampling path and upsampling path are three, nine, and one, respectively. Therefore, the resolution path expansion at each block provides the versatility to process feature maps with a combination of various receptive fields compared to the fixed single-scale branches. Indeed, we observed that the required resolution of branches varies with the length of utterances (see Fig. 2). Subsequently, upsampling and downsampling are applied to the low- and high-resolution paths, respectively, to restore the original resolution. This process encourages features to be handled in multiple time scales by selectively activating each branch based on utterance lengths.

At low-, original-, and high-resolution branches, individually calculated features $F^l(x)$,$F^o(x)$, and $F^h(x) \in \mathbb{R}^{T\times C}$ are as follows, where T and C denote the time and channel axes of the feature.

$$\begin{aligned} F^l(x) &= \sum_{i=1}^{8} U_i^l(f_i^l(D(x))), \\ F^o(x) &= \sum_{i=1}^{16} f_i^o(x), \\ F^h(x) &= \sum_{i=1}^{8} D(f_i^h(U_i^h(x))) \end{aligned} \tag{2}$$

where f_i^r is the convolutional layer of the i-th path in the r resolution branch, $r=\{l,o,h\}$. $D(x)$ is the downsampling function, which is an average pooling layer. $U_i^r(x)$ is the upsampling function, which is a transposed convolutional layer. Both have the same kernel and stride size of 3.

Furthermore, we introduce a multi-head attention-based gate module to dynamically fuse the activation of branches. Firstly, the features output from the three branches are averaged based on the time axis and then concatenated, denoted by $\mu(\cdot)$ and $[\cdot]$, respectively.

$$H = [\mu(F^l(x)),\mu(F^o(x)),\mu(F^h(x))],\ H \in \mathbb{R}^{3\times C} \tag{3}$$

$$W_t = Z^{\top}\sigma(Y^{\top}H_t + p) + q,\ W_t \in \mathbb{R}^{1\times C} \tag{4}$$

Afterward, the obtained vector H_t is transformed to the attention weights W_t through two linear layers (Y,p and Z,q), where t is the time axis. Then, the attention score, A_t^c is derived by applying the softmax operation to each channel of W, where c is the channel axis.

$$A_t^c = \frac{exp(W_t^c)}{\sum_{i=1}^{3} exp(W_i^c)},\ A_t^c \in \mathbb{R}^1 \tag{5}$$

Consequently, the multi-head attention-based gate module can selectively reflect the activation of each branch by multiplication between the feature and attention map, $A_t \in \mathbb{R}^{1\times C}$.

$$\begin{aligned} &Gate(F^l(x),F^o(x),F^h(x)) \\ &\quad = F^l(x)\odot A_1 + F^o(x)\odot A_2 + F^h(x)\odot A_3 \end{aligned} \tag{6}$$

$\odot$ refers to the element-wise multiplication after the attention map's time axis is broadcast. Finally, RawNeXt block with skip-path, B, is expressed as follows:

$$B(x) = \sigma(Gate(F^l(x),F^o(x),F^h(x)) + x) \tag{7}$$

The stack of RawNeXt blocks increases the combination of resolution path options exponentially, leading to dynamic propagation for variable-length utterances.

4. EXPERIMENTAL SETUP

4.1. Datasets

For training, we used the VoxCeleb2 dataset [15], which comprises over 1 million utterances from 6,112 speakers. To prove the effectiveness of our model under various conditions, we exploited three evaluation trials on the VoxCeleb1 dataset [14]. The original evaluation trial (Vox1-O) consists of 37,611 enrollment-test utterance pairs from 40 speakers, corresponding to the test set of the VoxCeleb1 dataset. The extended evaluation trial (Vox1-E) contains a list of 579,818 pairs from 1,251 speakers in the entire VoxCeleb1 dataset, and the hard evaluation trial (Vox1-H) includes a list of 550,894 pairs with the same nationality and gender from 1,190 speakers. We tested trials using cosine similarity and evaluated the models with the equal error rate (EER) and the minimum detection cost function (C_{det}^{min}), as in [15].

4.2. Implementation details

We employed the raw waveforms as input with pre-emphasis applied. For each iteration, the mini-batch consisted of 320 utterances (2 utterances from each of 160 randomly selected speakers). The lengths of the two utterances for each speaker were set to a fixed 59,049 samples and a random number of samples between 16,000 and 59,049. The random length utterances were duplicated to fit 59,049 samples for mini-batch construction. This configuration encourages the model to learn the EDSP strategy explicitly by training utterances of various lengths. In all experiments, we used the AMSGrad optimizer [28]. The initial learning rate (LR) was $1e^{-3}$ and decreased to $1e^{-7}$ for 80 epochs using a cosine LR scheduler. We set the weight decay to $1e^{-4}$. In several experiments, data augmentation was applied using room impulse response simulation and the MUSAN corpus [29].

Table 3: Comparison with VoxSRC-20's top-three single speaker verification systems [16] on the three different evaluation trials.

Model	Input Feature	Loss Function	Aggregation Method	Vox1-O	Vox1-E	Vox1-H
				EER% / C_{det}^{min}	EER% / C_{det}^{min}	EER% / C_{det}^{min}
ResNet-100m2 [25]	MFB-80	AM-Softmax	SP	1.1 / **0.064**	-	-
DPN68 [26]	MFB-40	CM-Softmax	SP	0.77 / 0.077	0.96 / 0.103	1.66 / **0.156**
ECAPA-TDNN [27]	MFCC-80	AAM-Softmax	ASP	**0.56** / 0.074	**0.84** / **0.096**	**1.57** / 0.164
RawNeXt	Waveform	AAM-Softmax	ASP	**1.29** / 0.142	**1.17** / **0.138**	2.28 / 0.236
RawNeXt	Waveform	AAM-Softmax+AP	ASP	1.32 / **0.136**	1.19 / 0.145	**2.23** / **0.228**

Table 4: Ablation experiments of RawNeXt components. (D: Deep layer aggregation, E: Elastic, U: upsampling path, G: Gate module)

Systems	D	E	G	U	Vox1-O (EER%)			
					1s	2s	5s	full
#1(ResNeXt)	×	×	×	×	6.12	3.68	2.45	2.16
#2	✓	×	×	×	4.82	2.98	2.08	1.93
#3	×	✓	×	×	5.39	3.18	2.16	1.95
#4	✓	✓	×	×	4.66	2.94	2.13	1.94
#5	✓	✓	✓	×	4.67	3.01	2.08	1.88
#6	✓	✓	×	✓	4.65	2.81	1.94	1.82
#7(RawNeXt)	✓	✓	✓	✓	**4.47**	**2.58**	**1.72**	**1.54**

5. RESULTS

In Table 2, we compare our model with the recently proposed systems for short utterances. To evaluate the performance with short utterances on the Vox1-O trial, we used full-duration enroll utterances and test utterances truncated to durations of 1, 2, and 5 seconds. The test utterance was cropped in the middle of the utterance, and if the utterance length was shorter than the target length, it was duplicated. As a result of the experiments, baseline ResNeXt showed relatively satisfactory performance for the full-duration test, with a better result than RawNet2 under the same conditions. However, significant performance degradation occurred for the variable-duration scenario compared to the recently proposed systems (rows 1,2,3). Proposed RawNeXt outperformed other models with different input features or improved loss functions under all test conditions. In addition, RawNeXt with the combination of data augmentation and AAM-softmax [30] loss achieved state-of-the-art results on short-utterance SV scenarios as well as full-duration utterances. Based on these results, we judge that the proposed model is effective for variable-duration utterances and can be enhanced by using various loss functions.

Table 3 shows the comparison with the top-three systems of VoxSRC-20, which reported state-of-the-art performance in SV. We trained RawNeXt using improved loss functions, such as AAM-softmax and angular prototypical network (AP) [31], and all experiments in this table used data augmentation. Although our system was proposed for variable-duration utterances, it exhibited relatively tolerable performance compared to the top-three systems. These results suggest that RawNeXt has high potential for not only short utterances but also generalization of SV.

Table 4 presents the results of ablation experiments to demonstrate the efficacy of each RawNeXt's component for variable-duration utterances. The results of Systems #1, 2, 3, and 4 show that the use of DLA and Elastic, proposed for computer vision tasks, leads to better SV performance. This implies that the motivations of each method are well aligned with the goal of short-utterance SV. A comparison of Systems #4, 5, 6, and 7 suggests that using both our proposed gate module and upsampling path (meaning EDSP) complements the original Elastic scheme, resulting in additional performance improvement for variable-duration utterances.

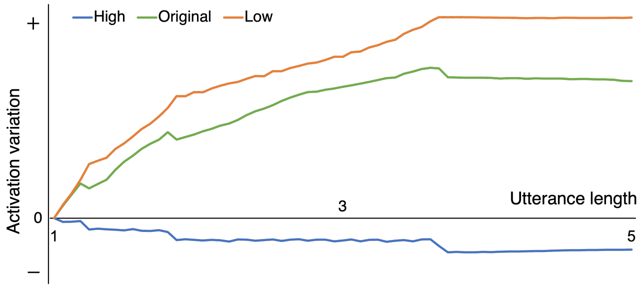

Fig. 2: Variation score for mean activation of each resolution path according to the input utterance length on VoxCeleb1 test set.

Furthermore, to analyze the trained EDSP of RawNeXt, we defined the score S_L^r at each r resolution branch by differences of mean activations between L and a 1-second utterance as follows:

$$S_L^r = \frac{1}{TC}\left(\sum_{t=1}^{T}\sum_{c=1}^{C} x_{L_{tc}}^r - \sum_{t=1}^{T}\sum_{c=1}^{C} x_{1_{tc}}^r\right) \tag{8}$$

where, T and C are the frame length and the number of channels of the feature, respectively. x_L^r is the tensor of L second utterance derived by multiplying the attention map with the activation of the r resolution branch, as in each term of eq. (6). Fig. 2 visualizes the activation variation score (average over all layers) of each resolution branch according to the utterance length. Obviously, as the length of the input utterance increases, it becomes less active in the high-resolution path and more active in the low-resolution path. This tendency implies that the model can extract speaker information with appropriate resolutions by dynamically applying scaling policies according to the length of the utterance, as discussed in section 3.2. Thus, short utterances containing relatively sparse speaker information require exquisite feature extraction with small receptive fields at higher resolutions, whereas long utterances require more comprehensive feature extraction with large receptive fields at lower resolutions.

6. CONCLUSION

We proposed a novel speaker verification (SV) system, RawNeXt, using a deep layer aggregation (DLA) structure and extended dynamic scaling polices (EDSP) for variable-duration input utterances. The DLA extracts speaker embeddings rich in time-spectral information by aggregating the time scales and spectral channels of features in the network, iteratively and hierarchically. The EDSP dynamically captures speaker-discriminative information by selectively activating the branches of different resolutions in blocks according to the length of utterances. As a result of the evaluation on the VoxCeleb dataset, RawNeXt outperformed the recently proposed baseline systems for shorter utterances and demonstrated a strong generalization ability of SV while exhibiting comparable performance to the state-of-the-art systems. In addition, we proved the effectiveness of our system's components through ablation and analysis experiments.

7. REFERENCES

[1] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

[2] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[3] 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.

[4] Chunlei Zhang and Kazuhito Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances.," in *Interspeech*, 2017, pp. 1487–1491.

[5] Jee-weon Jung, Hee-Soo Heo, Hye-jin Shim, and Ha-Jin Yu, "Short utterance compensation in speaker verification via cosine-based teacher-student learning of speaker embeddings," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 335–341.

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

[7] Zhifu Gao, Yan Song, Ian Vince 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 *INTERSPEECH*, 2019, pp. 361–365.

[8] Youngmoon Jung, Seong Min Kye, Yeunju Choi, Myunghun Jung, and Hoirin Kim, "Improving multi-scale aggregation using feature pyramid module for robust speaker verification of variable-duration utterances," *Proc. Interspeech 2020*, pp. 1501–1505, 2020.

[9] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell, "Deep layer aggregation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[10] Mirco Ravanelli and Yoshua Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028.

[11] Peng Zhang, Peng Hu, and Xueliang Zhang, "Deep embedding learning for text-dependent speaker verification.," in *INTERSPEECH*, 2020, pp. 3461–3465.

[12] Huiyu Wang, Aniruddha Kembhavi, Ali Farhadi, Alan L Yuille, and Mohammad Rastegari, "Elastic: Improving cnns with dynamic scaling policies," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2258–2267.

[13] Jee-weon Jung, Seung-bin Kim, Hye-jin Shim, Ju-ho Kim, and Ha-Jin Yu, "Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms," *Proc. Interspeech 2020*, pp. 3583–3587, 2020.

[14] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, pp. 101027, 2020.

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

[16] Arsha Nagrani, Joon Son Chung, Jaesung Huh, Andrew Brown, Ernesto Coto, Weidi Xie, Mitchell McLaren, Douglas A Reynolds, and Andrew Zisserman, "Voxsrc 2020: The second voxceleb speaker recognition challenge," *arXiv preprint arXiv:2012.06867*, 2020.

[17] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[18] 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.

[19] Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proc. Interspeech 2018*, 2018, pp. 2252–2256.

[20] Seong Min Kye, Joon Son Chung, and Hoirin Kim, "Supervised attention for speaker recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 286–293.

[21] Seong Min Kye, Youngmoon Jung, Hae Beom Lee, Sung Ju Hwang, and Hoirin Kim, "Meta-Learning for Short Utterance Speaker Recognition with Imbalance Length Pairs," in *Proc. Interspeech 2020*, 2020, pp. 2982–2986.

[22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[23] Yun Tang, Guo-Hong Ding, Jing Huang, Xiaodong He, and Bowen Zhou, "Deep speaker embedding learning with multi-level pooling for text-independent speaker verification," *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6116–6120, 2019.

[24] Haiying Zhao, Wei Zhou, Xiaogang Hou, and Hui Zhu, "Double attention for multi-label image classification," *IEEE Access*, vol. 8, pp. 225539–225550, 2020.

[25] Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck, "The idlab voxceleb speaker recognition challenge 2020 system description," *arXiv preprint arXiv:2010.12468*, 2020.

[26] Xu Xiang, "The xx205 system for the voxceleb speaker recognition challenge 2020," *arXiv preprint arXiv:2011.00200*, 2020.

[27] Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck, "The idlab voxsrc-20 submission: Large margin fine-tuning and quality-aware score calibration in dnn based speaker verification," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5814–5818.

[28] J REDDI Sashank, Kale Satyen, and Kumar Sanjiv, "On the convergence of adam and beyond," in *International Conference on Learning Representations*, 2018, vol. 5, p. 7.

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

[30] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[31] 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*, 2020, pp. 2977–2981.