



# Reshape Dimensions Network for Speaker Recognition

Ivan Yakovlev, Rostislav Makarov, Andrei Balykin, Pavel Malov, Anton Okhotnikov,  
Nikita Torgashov

ID R&D Inc., New York, USA

{yakovlev, makarov, andrew.balykin, pavel.malov, ohotnikov, torgashov}@idrnd.net

## Abstract

In this paper, we present **Reshape Dimensions Network** (ReDimNet), a novel neural network architecture for extracting utterance-level speaker representations. Our approach leverages dimensionality reshaping of 2D feature maps to 1D signal representation and vice versa, enabling the joint usage of 1D and 2D blocks. We propose an original network topology that preserves the volume of channel-timestep-frequency outputs of 1D and 2D blocks, facilitating efficient residual feature maps aggregation. Moreover, ReDimNet is efficiently scalable, and we introduce a range of model sizes, varying from 1 to 15 M parameters and from 0.5 to 20 GMACs. Our experimental results demonstrate that ReDimNet achieves state-of-the-art performance in speaker recognition while reducing computational complexity and the number of model parameters.

**Index Terms:** speaker recognition, speaker verification, speech processing, ReDimNet

## 1. Introduction

Speaker recognition is a specialized field aiming at identifying or verifying individuals through their distinct voice features. In this domain, deep neural networks have emerged as a major technology for extracting speaker embeddings that are used for multiple tasks including Speaker Verification (SV), Speaker Identification, Speaker Diarization, and others. Extensive research has been conducted in the SV area, which includes the development of new datasets [1–4], model architecture designing [5–18], and inventing new loss functions [19, 20].

A variety of architectures have emerged including 1D [5–7, 9, 10] and 2D [14–18] convolutional neural networks (CNNs), their hybrids that incorporate 2D CNN stem before 1D TDNN-like backbone [8, 11, 13], as well as self-attention networks [12]. Each architectural approach brings its unique set of advantages with 1D models offering efficiency and direct temporal analysis, 2D architectures providing frequency translational invariance [21], and hybrid systems aiming to deliver the best of both worlds. Additionally, design approaches can be split into macro and micro designs, with micro designs involving modifications like substituting traditional 1D ResBlocks with Res2Net blocks within the ECAPA-TDNN architecture [6], and macro designs incorporating a 2D stem ahead of TDNN-like models [8, 11, 13] leading to a two-stage architecture that transitions 2D  $\rightarrow$  1D.

In this paper, we introduce ReDimNet<sup>1</sup>, a novel neural network architecture based on the dimensionality reshaping of feature maps between 2D and 1D representations, enabling seamless integration of 1D and 2D blocks. ReDimNet exhibits scalability across various model sizes, while consistently achiev-

<sup>1</sup><https://github.com/IDRnD/ReDimNet>

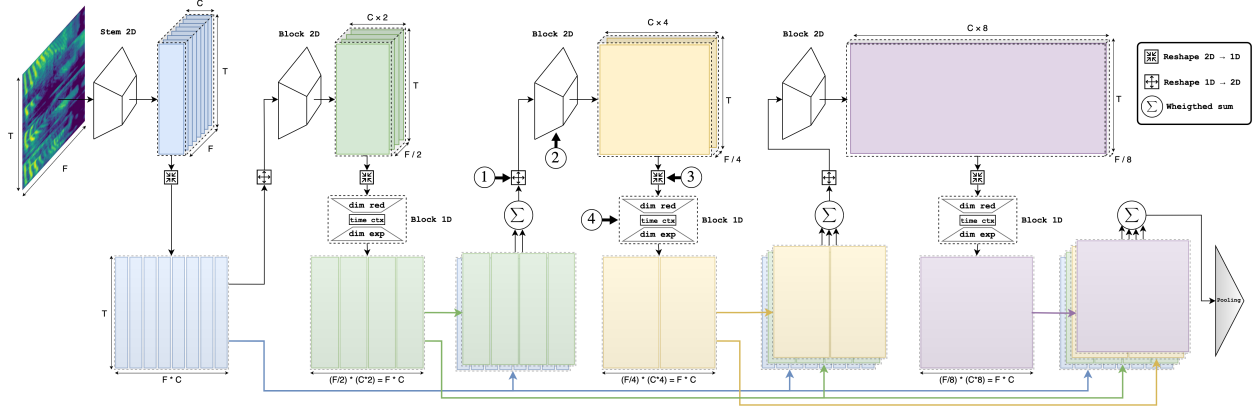


**Fig. 1 Computational Cost vs. Average Equal Error Rate.** EER is averaged over three Voxceleb1 protocols: Vox1-O, Vox1-E, Vox1-H. The model size is shown by the area of a circle, model family is indicated by a color. Complexity is assessed using `thop` library with an input signal of 2 seconds. A short dashed line represents scaling the ReDimNet architecture using the Additive Angular Margin loss function [19], dashed line - using the SphereFace2 loss [20].

ing optimal performance under varying computational resource constraints. Our experimental results demonstrate that ReDimNet outperforms many other architectures and achieves state-of-the-art performance on public benchmarks while reducing inference time and model size.

## 2. Model Architecture

In this section, we detail the design of the proposed architecture influenced by two main concepts. Firstly, to leverage the benefits of residual connections, we incorporate them extensively in ReDimNet. Secondly, based on the success of models utilizing both 1D and 2D blocks for speech processing and SV, our architecture integrates both types of blocks to boost performance.



**Fig. 2 ReDimNet architecture scheme.** Digits 1,2,3 and 4 describe the order of operators and blocks execution in a single model stage, where  $C$  - number of channels,  $F$  - number of frequency bins,  $T$  - number of timestamps.

### 2.1. Dimensions reshape & residual connection

The main distinguishable feature of the architecture is an ability to aggregate 1D and 2D feature maps together with 2D feature maps from other model stages to enable 1D-2D and 2D-2D skip connections with various feature maps shapes. Such a technique is implemented only using the reshape operation without broadcasting or dimensionality reduction/expansion. We constrain ReDimNet to output feature maps with predefined shapes, that are easily reshaped back and forth between fixed 1D representation and various sets of 2D representations. First, we mitigate all strides across the time axis in a model, meaning that in the end, before pooling, the model will have the same time resolution as input features. Second, we synchronize strides over frequency dimension with a growth rate of channels for all stages, to make a "volume" of 2D feature maps constant throughout model forward pass. Given that, if all 2D feature maps are represented by the tensors with common PyTorch [22] size notation ( $s_0 = batch\_size : bs, s_1 = channels : C, s_2 = frequency : F, s_3 = time : T$ ), we assign volume of 2D feature map as  $V = s_1 \cdot s_2 \cdot s_3$ . This property of ReDimNet architecture is well illustrated in the model scheme (Fig. 2) and in the Table 1 presenting the internal feature map size for each block.

**Table 1 Model internal feature map sizes.**  $S_f$  stands for frequency stride.

Block #	In shape	$S_f$	Channels	Out shape	Volume
1	$(C, F, T)$	1	$C$	$(C, F, T)$	$C \cdot F \cdot T$
2	$(C, F, T)$	2	$C \cdot 2$	$(C \cdot 2, F/2, T)$	
3	$(C \cdot 2, F/2, T)$	2	$C \cdot 4$	$(C \cdot 4, F/4, T)$	
4	$(C \cdot 4, F/4, T)$	2	$C \cdot 8$	$(C \cdot 8, F/8, T)$	
5	$(C \cdot 8, F/8, T)$	1	$C \cdot 8$	$(C \cdot 8, F/8, T)$	

Having the same volume in all 2D feature maps is not yet enough to sum them right away to enable skip connections due to the shape mismatch. However, this can be easily overcome by an introduction of invertible reshape operator that reshapes all 2D feature maps of size  $(bs, C_i, F_i, T)$  into 1D feature map of constant size:  $(bs, C_i \cdot F_i, T) = (bs, C_0 \cdot F_0, T)$ , where  $C_0 = C$  and  $F_0 = F$ . This equality is constant for various stages outputs due to the model strides and channels growth constraints. Then we sum 1D feature maps and reshape them back to 2D using the inverse reshape operator, this way we enable residual connection through the whole model forward pass.

### 2.2. 1D & 2D Blocks

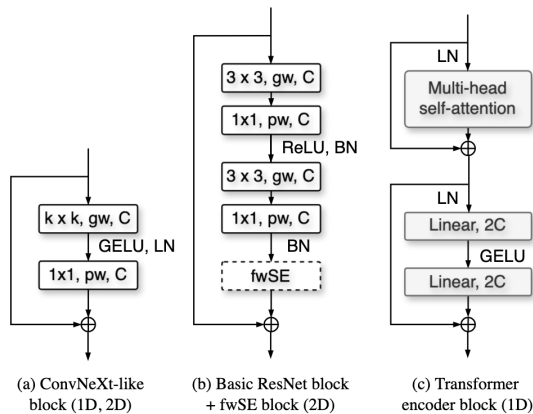
ReDimNet is created around the use of joint 1D and 2D blocks, which are presented correspondingly by Block 1D and Block 2D in the scheme in Fig. 2. These blocks are designed to handle 1D feature maps of fixed size, which are then reshaped into 2D feature maps for processing within the block. The structure of these blocks makes possible dynamic interchange between 2D and 1D representations: 2D subblocks process ② the reshaped (in ①) 1D inputs using sequences of residual blocks with 2D convolutions, and then the output is converted back to a 1D format ③ for further processing in the 1D subblock ④. This 1D subblock employs a channel-axis dimensionality reduction Fully Connected (FC) layer + normalization layer, followed by a time-contextual processing component. This component can be implemented through ConvNeXt-like 1D blocks, transformer encoder blocks, or a combination of them, and its output is a 1D feature map. Finally, the channel-axis expansion FC layer unfolds the number of channels to match input shape and performs skip+residual sum operation. More information on the basic blocks structure used in ReDimNet is provided in Fig. 3.

### 2.3. Input features & pooling

As model input features we used 72-dimensional mean-normalized Mel filter bank log-energies with a 25 ms frame length and 15 ms step with 512 FFT size over the 20-7600 Hz frequency range by default. To extract an utterance-level embedding from the frame-level features, we utilized the Attentive Statistics Pooling [23] with global context.

## 3. Experimental Setup

We conducted experiments of training ReDimNet architecture utilizing the development part of the VoxCeleb2 [1] dataset. Models were optimized using SGD optimizer with Nesterov momentum,  $m = 0.9$ , and a weight decay of  $2e^{-5}$ . As a default loss function, we selected Additive Angular Margin (AAM) softmax loss [19] due to its wide adoption. We also conducted and reported results for a few experiments with SphereFace2 (SF2) loss function [20] for comparison purposes. We followed a 2-stage training approach by firstly pretraining a model on short segments with multiple augmentations applied. Then, we



**Fig. 3 Block design.** In ReDimNet as 2D blocks we used (a) slightly modified ConvNeXt-like block [29] or (b) basic ResNet block [30] with fwSE [21]. As 1D blocks we used same (a) 1D version of ConvNeXt-like block with or in place of (c) Transformer block [31].

applied finetuning on longer utterances with some augmentations turned off and tweaked the parameters of a loss function. This second training stage is well-known as Large-Margin (LM) finetuning strategy [24]. All models were trained using the *wespeaker* [25] training pipeline.

### 3.1. Pretraining stage

For pretraining, we used a default *voxceleb2* recipe from *wespeaker* pipeline with minor adjustments. 2-second segments were selected randomly from each signal, and various augmentations with MUSAN dataset [26] (noise, music, babble) alongside the RIR dataset [27] were applied following the augmentation recipe from [14]. A two-fold speed augmentation [28], with factors of 0.9 and 1.1, was employed to generate additional speakers within the training dataset. In this stage, the AAM-softmax margin penalty was scheduled as follows: first 20 epochs it was kept at 0.0, then for the next 20 epochs it exponentially rose to 0.2 and then was kept constant till the end of training. We used Exponential Decay with Warmup learning rate scheduler with 6 epochs warmup,  $lr_{max} = 1e^{-1}$  and  $lr_{min} = 1e^{-5}$ .

### 3.2. Large-Margin Finetuning stage

At the finetuning stage [24], AAM-softmax margin was set to constant 0.5 value, with length of training utterances expanded to 6 seconds. Speed perturbations were turned off during this stage.

### 3.3. Evaluation

The performance of models is assessed using cleaned protocols of VoxCeleb1 [32] test set, employing the Equal Error Rate (EER) and the minimum Detection Cost Function (minDCF) with  $P_{target} = 0.01$  and  $C_{FA} = C_{Miss} = 1$ . We scored each model with cosine backend utilizing full utterance length as input and additionally applied a top-300 adaptive s-normalization (AS-Norm) [33] of cosine scores (see Table 4).

## 4. Model scaling & ablation studies

### 4.1. Model scaling

Achieving efficient model scaling was one of our main research goals. Therefore, we were able to scale the ReDimNet architecture from (1M, 0.5 GMACs) to (15M, 20 GMACs), reaching competitive results for each model size. For the naming convention, we followed notations of [34], resulting in 7 configurations: B0 - B6, where each model configuration is bounded by the computational complexity limits in GMACs that we found to be a predominant factor of model scaling relative to model size. Complete testing results of each ReDimNet configuration are shown in Table 4.

**Table 2** Ablation Study on Block Components of ReDimNet (EER, %)

	Block Type	Vox1-O	Vox1-E	Vox1-H	Average
1D block	Skip Connection	1.59	1.57	2.71	1.96
	Fully Connected	0.93	1.13	1.94	1.33
	1D Conv	0.65	0.85	1.54	1.01
	MHA	0.69	0.82	<b>1.45</b>	0.99
	1D Conv + MHA	<b>0.59</b>	<b>0.79</b>	1.47	<b>0.95</b>
	ConvNext block	0.68	0.83	<b>1.46</b>	0.99
2D block	fwse-ResNet block	0.64	0.82	1.48	0.98
	ResNet block	<b>0.61</b>	<b>0.80</b>	1.48	<b>0.96</b>

### 4.2. Ablation studies

We also conducted a thorough study of how different components of ReDimNet architecture affect its performance. This research includes studying the role of 1D and 2D blocks for speech signal processing, assessing the impact of different loss functions, and optimizing group sizes and steps in convolutions for accuracy and efficiency improvement. All ablation studies were carried out on the ReDimNet-B2 architecture.

**Table 3** Ablation study on loss function configuration (EER,%)

Loss Type	Vox1-O	Vox1-E	Vox1-H	Average
AAM-SC	<b>0.57</b>	0.91	1.60	1.03
AAM	0.68	0.83	1.46	0.99
SF2-A	0.63	0.80	1.39	0.94
SF2-C	<b>0.57</b>	<b>0.76</b>	<b>1.32</b>	<b>0.88</b>

#### 4.2.1. Block types

In order to identify optimal configurations of ReDimNet architecture, we compared three types of 2D-blocks: basic ResNet block, basic ResNet FWSE block, and a ConvNext block. While minimal differences were observed, basic ResNet block, however, slightly outperformed others by a small margin (see Table 2).

Our further analysis was focused on the 1D block type, where we assessed a range of options including sequences of 1D convolutional ConvNeXt-like blocks (Fig. 3), multi-head attention (MHA) (Fig. 3), FC layers, skip connections, and a hybrid of 1D convolutional blocks with MHA (1D Conv + MHA). Skip connections appeared to be the least effective approach, which underscored the importance of the 1D block within the ReDimNet architecture. FC layers performed slightly better, suggesting the importance of a temporal context. 1D convolutional and MHA blocks have proven to be the most efficient configurations, and a combination of MHA and 1D convolutional blocks delivered the best performance (see Table 2).

**Table 4 Evaluation results on the VoxCeleb1-Cleaned protocols without QMFs.** For the report, we calculated the equal error rate (EER) and the minimum detection cost function (minDCF). GMACs were measured on 2-s long segments. \* - means values have been estimated. Open source models from the WeSpeaker or ECAPA2 repositories were retested in our environment.

Model	Params	GMACs	LM	AS-Norm	Vox1-O		Vox1-E		Vox1-H	
					EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
<b>ReDimNet-B0</b>			✓	✗	1.16	0.101	1.25	0.132	2.20	0.207
+AS-Norm	<b>1.0M</b>	<b>0.43</b>	✓	✓	<b>1.07</b>	<b>0.098</b>	<b>1.18</b>	<b>0.121</b>	<b>2.01</b>	<b>0.184</b>
NeXt-TDNN-1 (C=128, B=3) [7]	<b>1.6M</b>	<b>0.29*</b>	✗	✓	1.10	0.108	1.24	0.133	2.12	0.201
NeXt-TDNN (C=128, B=3) [7]	1.9M	0.35*	✗	✓	1.03	0.095	1.17	0.126	1.98	0.190
<b>ReDimNet-B1</b>			✓	✗	0.85	0.076	0.97	0.104	1.73	0.166
+AS-Norm	2.2M	0.54	✓	✓	<b>0.73</b>	<b>0.071</b>	<b>0.89</b>	<b>0.096</b>	<b>1.57</b>	<b>0.154</b>
ECAPA (C=512) [6, 17]	6.4M	1.05	✗	✓	0.94	0.092	1.21	0.129	2.20	0.205
NeXt-TDNN-1 (C=256, B=3) [7]	6.0M	1.13*	✗	✓	0.81	0.091	1.04	0.116	1.86	0.184
CAM++ [11, 25]	7.2M	1.15	✓	✗	0.71	0.109	0.85	0.095	1.66	0.165
NeXt-TDNN (C=256, B=3) [7]	7.1M	1.35*	✗	✓	0.79	0.087	1.04	0.115	1.82	0.182
<b>ReDimNet-B2<sub>SF2</sub></b>			✓	✗	0.57	0.054	0.76	0.082	1.32	0.133
+AS-Norm	<b>4.7M</b>	<b>0.90</b>	✓	✓	<b>0.52</b>	<b>0.060</b>	<b>0.74</b>	<b>0.078</b>	<b>1.27</b>	<b>0.128</b>
ECAPA (C=1024) [6, 17]	14.9M	2.67	✓	✗	0.98	0.105	1.13	0.117	2.09	0.204
DF-ResNet56 [17]	4.5M	2.66	✗	✓	0.96	0.103	1.09	0.122	1.99	0.184
Gemini DF-ResNet60 [18]	4.1M	<b>2.50*</b>	✗	✓	0.94	0.089	1.05	0.116	1.80	0.166
<b>ReDimNet-B3</b>			✓	✗	0.50	0.063	0.73	0.079	1.33	0.135
+AS-Norm	<b>3.0M</b>	3.00	✓	✓	<b>0.47</b>	<b>0.042</b>	<b>0.69</b>	<b>0.072</b>	<b>1.23</b>	<b>0.121</b>
ResNet34 [25, 30]	6.6M	<b>4.55</b>	✓	✗	0.82	0.080	0.93	0.104	1.68	0.161
Gemini DF-ResNet114 [18]	6.5M	5.00	✗	✓	0.69	0.067	0.86	0.097	1.49	0.144
<b>ReDimNet-B4</b>			✓	✗	0.51	0.052	0.68	0.073	1.26	0.123
+AS-Norm	<b>6.3M</b>	4.80	✓	✓	<b>0.44</b>	<b>0.042</b>	<b>0.64</b>	<b>0.067</b>	<b>1.17</b>	<b>0.111</b>
Gemini DF-ResNet183 [18]	9.2M	<b>8.25</b>	✗	✓	0.60	0.064	0.81	0.090	1.44	0.137
DF-ResNet233 [17]	12.3M	11.17	✗	✓	0.58	0.044	0.76	0.083	1.44	0.146
<b>ReDimNet-B5<sub>SF2</sub></b>			✓	✗	0.43	0.039	0.61	0.062	1.08	0.102
+AS-Norm	<b>9.2M</b>	9.87	✓	✓	<b>0.39</b>	<b>0.037</b>	<b>0.59</b>	<b>0.057</b>	<b>1.05</b>	<b>0.095</b>
ResNet293 [25, 30]	23.8M	28.10	✓	✗	0.53	0.057	0.71	0.072	1.30	0.127
ECAPA2 [13]	27.1M	187.00*	✓	✗	0.44	0.041	0.62	0.066	1.15	0.114
<b>ReDimNet-B6<sub>SF2</sub></b>			✓	✗	0.40	0.033	0.55	0.052	1.05	0.104
+AS-Norm	<b>15.0M</b>	<b>20.27</b>	✓	✓	<b>0.37</b>	<b>0.030</b>	<b>0.53</b>	<b>0.051</b>	<b>1.00</b>	<b>0.097</b>

#### 4.2.2. Loss studies

Furthermore, we explored the effectiveness of various loss functions (see Table 3). Specifically, we evaluated SphereFace losses (SF2) with A and C configurations [20], Additive Angular Margin Loss (AAM), and Additive Angular Margin loss with SubCenters (AAM-SC) [19]. Based on the testing results, we found SphereFace type C to be the most effective loss function providing the largest performance improvement in the benchmarks.

## 5. Results

Testing results of all proposed ReDimNet architecture configurations are presented in Table 4. We compared ReDimNet on the VoxCeleb1 protocols with publicly available models and grouped them based on the number of parameters and multiply-accumulate operations (MACs) for comparison purposes.

In particular, our ReDimNet-B1 model achieves comparable results to NeXt-TDNN [7] on the Vox1-H protocol, but has a slightly larger number of parameters and MACs. ReDimNet-B3 outperforms Gemini DF-ResNet60 [18] and ECAPA (C = 1024) with an advantage in model size. ReDimNet-B5 further improves upon the B3 version, consistently achieving the lowest EER and minDCF, compared to DF-ResNet233 [17], which has the similar number of parameters and MACs. Moreover, our largest model, ReDimNet-B6, delivers even better results while having significantly fewer parameters and MACs than ECAPA 2 [13] and ResNet293 [25, 30].

Furthermore, we subjected the best models of various architectures to additional out-of-domain testing (see Table 5). These results demonstrate that ReDimNet-B6 outperforms other architectures with a significant gap on unseen data domains.

**Table 5** Evaluation results on Speakers In The Wild core-core protocol [35], VoICES from a Distance Challenge Evaluation Set [36] and VoxCeleb1-B protocol [37] (EER, %).

Model	SITW	VoICES	Vox1-B	Average
CAM++	1.34	6.30	2.79	3.48
ECAPA (C=1024)	1.67	5.31	3.48	3.49
ResNet293	1.67	5.14	2.23	3.01
ECAPA2	3.64	13.26	1.81	6.24
<b>ReDimNet-B6</b>	<b>0.77</b>	<b>3.19</b>	<b>1.66</b>	<b>1.87</b>

## 6. Conclusions

In this paper we introduced ReDimNet - a novel neural network architecture designed for the extraction of utterance-level speaker representations. It combines dimensionality reshaping, dynamic transitions between 1D and 2D representations, and 2D and 1D blocks. Through a comprehensive evaluation, ReDimNet demonstrated:

- architecture adaptability and scalability across multiple configurations;
- top balance between computational efficiency and performance;
- strong results on the VoxCeleb1-H (cleaned) protocol, with an Equal Error Rate (EER) of 1.00%;
- advanced generalization ability on out-of-domain test sets.

In summary, ReDimNet architecture achieves competitive performance on all tests compared to other state-of-the-art speaker recognition models, while also offering favorable computational efficiency. Its adaptability and superior performance make it a valuable contribution to the speaker recognition field and a promising solution for real-world applications.

## 7. References

- [1] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech 2018*. ISCA, Sep. 2018.
- [2] Y. Lin, X. Qin, G. Zhao, M. Cheng, N. Jiang, H. Wu, and M. Li, "Voxblink: A large scale speaker verification dataset on camera," 2023.
- [3] S. Zheng, L. Cheng, Y. Chen, H. Wang, and Q. Chen, "3d-speaker: A large-scale multi-device, multi-distance, and multi-dialect corpus for speech representation disentanglement," 2023.
- [4] I. Yakovlev, A. Okhotnikov, N. Torgashov, R. Makarov, Y. Vo-evodin, and K. Simonchik, "VoxTube: a multilingual speaker recognition dataset," in *Proc. INTERSPEECH 2023*, 2023, pp. 2238–2242.
- [5] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.
- [6] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.
- [7] H.-J. Heo, U.-H. Shin, R. Lee, Y. Cheon, and H.-M. Park, "Next-tdnn: Modernizing multi-scale temporal convolution backbone for speaker verification," 2023.
- [8] T. Liu, R. K. Das, K. A. Lee, and H. Li, "Mfa: Tdnn with multi-scale frequency-channel attention for text-independent speaker verification with short utterances," 2022.
- [9] Z. Zhao, Z. Li, W. Wang, and P. Zhang, "Pcf: Ecapa-tdnn with progressive channel fusion for speaker verification," 2023.
- [10] Y.-Q. Yu and W.-J. Li, "Densely connected time delay neural network for speaker verification," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 921–925.
- [11] H. Wang, S. Zheng, Y. Chen, L. Cheng, and Q. Chen, "CAM++: A Fast and Efficient Network for Speaker Verification Using Context-Aware Masking," in *Proc. INTERSPEECH 2023*, 2023, pp. 5301–5305.
- [12] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, Z. Wu, H. yi Lee, and H. Meng, "Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification," 2022.
- [13] J. Thienpondt and K. Demuynck, "Ecapa2: A hybrid neural network architecture and training strategy for robust speaker embeddings," 2024.
- [14] D. Garcia-Romero, G. Sell, and A. Mccree, "Magneto: X-vector magnitude estimation network plus offset for improved speaker recognition," in *Odyssey*, 2020, pp. 1–8.
- [15] T. Zhou, Y. Zhao, and J. Wu, "Resnext and res2net structures for speaker verification," 2020.
- [16] Y. Chen, S. Zheng, H. Wang, L. Cheng, Q. Chen, and J. Qi, "An Enhanced Res2Net with Local and Global Feature Fusion for Speaker Verification," in *Proc. INTERSPEECH 2023*, 2023, pp. 2228–2232.
- [17] B. Liu, Z. Chen, S. Wang, H. Wang, B. Han, and Y. Qian, "DF-ResNet: Boosting Speaker Verification Performance with Depth-First Design," in *Proc. Interspeech 2022*, 2022, pp. 296–300.
- [18] T. Liu, K. A. Lee, Q. Wang, and H. Li, "Golden gemini is all you need: Finding the sweet spots for speaker verification," 2023.
- [19] J. Deng, J. Guo, N. Xue, and S. 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.
- [20] B. Han, Z. Chen, and Y. Qian, "Exploring binary classification loss for speaker verification," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [21] J. Thienpondt, B. Desplanques, and K. Demuynck, "Integrating frequency translational invariance in tdnn and frequency positional information in 2d resnets to enhance speaker verification," *arXiv preprint arXiv:2104.02370*, 2021.
- [22] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.
- [23] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Interspeech 2018*. ISCA, Sep. 2018.
- [24] L. H. Y. Liu and J. Liu, "Large margin softmax loss for speaker verification," in *in INTERSPEECH, 2019*, 2023, pp. 2873–2877.
- [25] H. Wang, C. Liang, S. Wang, Z. Chen, B. Zhang, X. Xiang, Y. Deng, and Y. Qian, "Wespeaker: A research and production oriented speaker embedding learning toolkit," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [26] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.
- [27] I. Szöke, M. Skácel, L. Mošner, J. Paliesek, and J. Černocký, "Building and evaluation of a real room impulse response dataset," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 863–876, 2019.
- [28] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Interspeech*, vol. 2015, 2015, p. 3586.
- [29] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," 2022.
- [30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.
- [32] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2019.
- [33] S.-C. Yin, R. Rose, and P. Kenny, "Adaptive score normalization for progressive model adaptation in text independent speaker verification," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4857–4860.
- [34] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2020.
- [35] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (sitw) speaker recognition database," in *Interspeech*, 2016, pp. 818–822.
- [36] M. K. Nandwana, J. van Hout, M. McLaren, C. Richey, A. Lawson, and M. A. Barrios, "The voices from a distance challenge 2019 evaluation plan," 2019.
- [37] K. Nam, Y. Kim, J. Huh, H. S. Heo, J. weon Jung, and J. S. Chung, "Disentangled representation learning for multilingual speaker recognition," 2023.