RCT-Net: TDNN based Speaker Verification with 2D Res2Nets on Frame Level Feature Extractor

*Note: Sub-titles are not captured in Xplore and should not be used

Razieh Khamsehashari
Quality and Usability
Technical University of Berlin
Berlin, Germany
email: razieh.khamsehashari@tu-berlin.de

Fengying Miao
Quality and Usability
Technical University of Berlin
Berlin, Germany
email: fengying.miao@campus.tu-berlin.de

Tim Polzehl
Speech and Language Technology
German Research Center for Artificial Intelligence (DFKI)
Berlin, Germany
email: tim.polzehl@dfki.de

Sebastian Möller
Quality and Usability
Technical University of Berlin
Berlin, Germany
email: sebastian.moeller@tu-berlin.de

***Abstract*—In speaker verification, Time Delay Neural Networks (TDNNs) and Residual Networks (ResNets) are currently achieving cutting-edge results. These architectures have very different structural characteristics, and development of hybrid networks appears to be a promising path forward. In this study, inspired by the combination of Convolutional Neural Network (CNN) blocks and multi-scale architectures we present a Residual-based CNN TDNN (RCT) system and evaluate the performance of integrating different residual blocks into a TDNN-based structure. We extend the state-of-the-art speaker embedding model for speaker recognition, namely Emphasized Channel Attention, Propagation, and Aggregation based CNN-TDNN (ECAPA CNN-TDNN), by gradually incorporating the proposed 2D convolutional stem with various bottleneck residual blocks. We evaluate the performance of our models on standard VoxCeleb1-O test set to investigate the performance of residual blocks and TDNN in the speaker verification domain. As a result, the proposed models significantly outperform the state-of-the-art by up to** 14.6% **of EER.**

***Index Terms*—ResNet, Residual blocks, TDNN, RCT-Net, speaker verification, automatic speaker verification (ASV)**

I. Introduction

Current state-of-the-art speaker verification systems try to improve the most popular neural network topology based on ECAPA-TDNN by incorporating multiple ideas and techniques inspired by convolutional blocks, feature aggregation, and frequency-channel attention methods. ECAPA CNN-TDNN [6] introduced a 2D convolutional stem for the ECAPA-TDNN, incorporating frequency translational invariance in the four top layers of the network. Liu et al. [7] proposed MFA-TDNN, a Multi-scale Frequency-channel Attention (MFA) framework, that captures the local information and frame-level temporal information by the dual-pathway multi-scale module while emphasizing the important frequency and channel components in TDNN systems. Inspired by ECAPA CNN-TDNN, which enhances ECAPA-TDNN by incorporating a CNN-based front-end, the MFA module is created as a front-end module for TDNNs in order to learn multi-scale and extract high resolution feature representations from short utterances. [8] and [13] adapt the frame-level processing in ECAPA-TDNN. In [8], their experiments focus on bottleneck residual blocks, attention mechanisms, and feature aggregation based on ECAPA-TDNN. They replaced the Res2Block with SC-Block and proposed the hierarchical feature aggregation method to build their final model.

Many recent studies have focused on expanding the receptive field of the convolutional layer on Residual Network (ResNet) [1]. The first technique integrates the ResNet with the concept of inception [2] and proposes ResNext, a split-transform-merge strategy [3]. The introduced *cardinality* is intended for processing different sizes of receptive fields in order to obtain multi-scale features. Furthermore, Res2Net [4] improves multi-scale feature extraction capability by constructing hierarchical residual-like connections within one single residual block. The preceding ideas are similar to the TDNN, which obtains a wide range of time information through convolution with different dilation rates. We believe that development of hybrid networks to generate multi-scale features influences the final representation and appears to be a promising direction moving forward. The ECAPA-TDNN model [5], as an example, combines the benefits of Res2Net and TDNN.

Inspired by these recent progresses, we propose Residual-based CNN TDNN *RCT-Net* using 2D convolutions based on different residual blocks as the foundation for the initial network layers. We evaluate the performance of various residual blocks using the most recent speaker embedding model for

ISBN: 978-1-68558-057-5

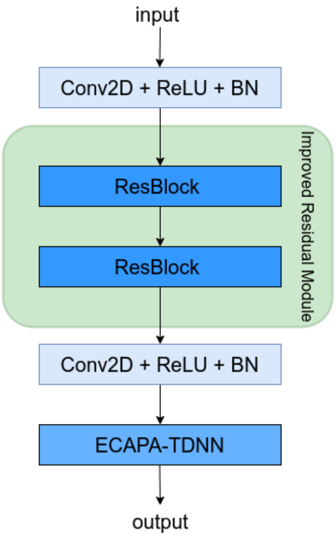


Fig. 1. The diagram of the proposed architecture.

speaker recognition, ECAPA CNN-TDNN [6], and experiment with the proposed 2D convolutional stem, including various bottleneck residual blocks such as Res2Net [4], Res2NeXt [3], standard ResNet [1], Improved ResNet [9] and ResTCN [10], [15].

This paper is organized as follows: In Section II, the baseline architectures are described. The structure of the proposed Residual-based CNN TDNN *RCT-Net* and different frame-level architectures are described in Section III. Section IV introduces the experimental setup including dataset, training the speaker embedding extractors, and evaluation protocol. Results and analysis are presented in Section V. In Section VI we discuss the potential justification for our best combination of two strong structures of TDNN and residual blocks. Finally, Section VII summarizes the findings.

II. Baseline system architectures

Two types of TDNN-based speaker embedding models are considered as reliable baselines to evaluate the performance of our suggested architecture: ECAPA-TDNN and ECAPA CNN-TDNN, which both currently provide state-of-the-art on speaker verification tasks.

The ECAPA-TDNN [5] model, which is based on the x-vector architecture [11], attempts to obtain exceptionally accurate x-vectors by introducing a number of enhancements to provide more robust speaker embeddings. First, channel- and context-dependent statistics pooling layer is used to aggregate all frame-level features to generate a fixed dimensional vector. Second, in order to add global context information to the locally operating convolutional blocks, the 1-dimensional Squeeze-Excitation (SE) block [17] is used and integrated with Res2Block [4], which has the advantage of multi-scale feature processing through group convolutions in hierarchical residual connections, and reduces the number of network parameters. Finally, the output features of all the SE-Res2Block for each frame are concatenated by multi-layer feature aggregation technique.

Inspired by 2D-CNNs, Thienpondt et al. [6] introduced a 2D convolutional stem in ECAPA-TDNN to transfer the advantages of ResNet architecture to the proposed hybrid CNN-TDNN network. Using ResNet in top layers allows the network to initially construct local, frequency-invariant features and then 1D convolutions are applied to incorporate the frequency position information of the features. The flattened output feature map subsequently is used to feed the ECAPA-TDNN network.

III. Proposed RCT-Net Architecture

The neural network is used by the current speaker verification methods to derive speaker representations. The effective x-vector architecture [11] uses TDNN to project variable-length utterances into fixed-length speaker characterization embeddings by applying statistics pooling. On the task of speaker verification, we aim to obtain an extremely accurate version of x-vector topology and try to enhance the performance of the original TDNN-based architectures [12].

We investigate different deep residual unit variations, and we are particularly interested in whether the TDNN and the basic residual building blocks simplicity can be successfully combined with the advantages of standard residual-based architectures [1] [9] [10] [14], and how the performance of the resulting architectures compares to the more sophisticated multi-scale residual blocks [3] [4]. In this regard, our method integrates, extends, and generalizes the architecture of ASV we previously described [13]. The proposed architecture, as shown in Figure 1, follows an established multi-scale and frequency positional encoding structure, ECAPA CNN-TDNN. In this study, we propose enhancements to the frame-level feature extractor.

A. Standard Residual Blocks

We shortly go over the key concepts underlying residual-based architectures like ResNet and Res-TCN [10] [15]. ResNet employs injected residual connections between processing streams to allow spatial-temporal interaction between them. Res-TCN redesigned the original TCN [14] by factoring out the deeper layers into additive residual terms that yielded both an interpretable hidden representation and model parameters. In contrast to the original ResNet, the basic residual unit of Res-TCN and improved ResNet [9] does not use ReLUs to support the element-wise additions $\bigoplus$ (see Figure 2(a-c)) and can therefore offer representations that are more interpretable. Additionally, such units create a direct path that enables the gradients and the signal to be transmitted directly in a backward pass through the entire network to any unit.

B. Multi-Scale Residual Blocks

Multi-scale feature representation has been integrated from the beginning into the CNN architectural design with a stack

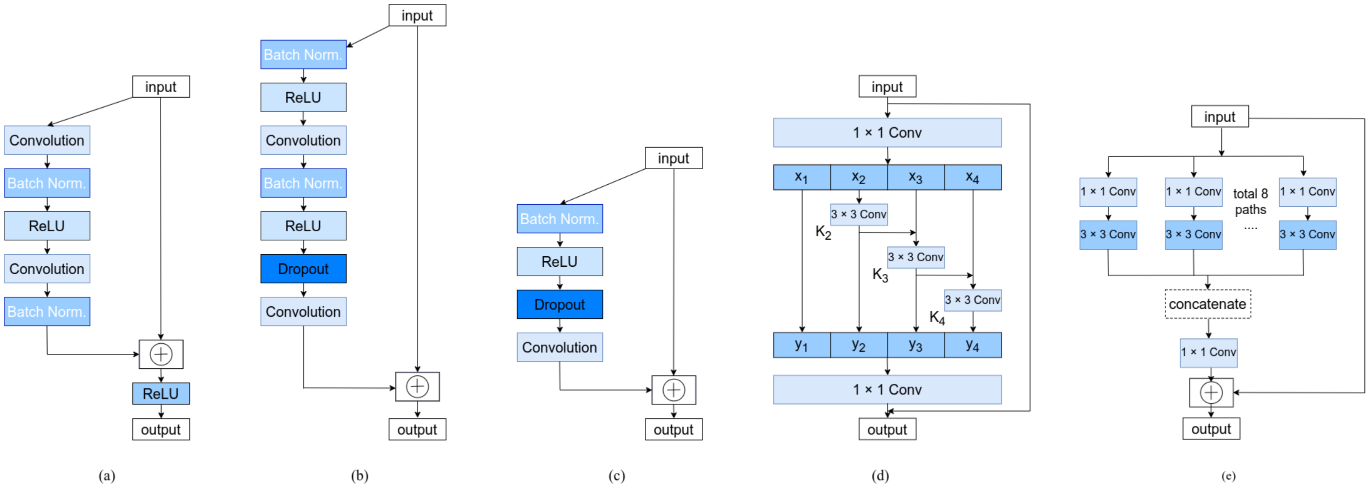


Fig. 2. The structures of bottleneck residual blocks in different architectures. Standard residual blocks in (a) ResNet [1], (b) Improved ResNet [9], and (c) Res-TCN [15]. Multi-scale residual blocks in (d) Res2Net [4] and (e) ResNeXt. [3]

of convolutional layers that automatically learn coarse-to-fine features [16]. The bottleneck module and shortcut connections to residual networks are effective at reducing the number of parameters and successfully addressing the gradient disappearance in deep CNN designs.

ResNeXt-50 [3] enhanced the bottleneck module by adding cardinal dimension and replacing conventional convolution with group convolution to perform more sophisticated transformations. Gao et al. [4] substituted the 3×3 convolution with a series of 3×3 convolution with smaller filter groups that are coupled hierarchically in order to incorporate the multi-scale capability of the feature representation into the module. This might be considered a network inside of a network. As a result, the range of receptive fields for each network layer is increased by the Res2NeXt, which also represents multi-scale features at a finer level. Res2NeXt-50 improved ResNeXt-50 by enabling multi-scale feature representation at both the global and local levels by integrating hierarchical multi-scale feature representation into the bottleneck module. SE-Res2NeXt-50 [4] integrated the SE block [17] to provide a channel-wise dynamic calibration of feature responses and provide enhanced feature representation capabilities.

Res2NeXt substitutes a set of 3×3 filters with smaller groups of filters, while connecting different filter groups in a hierarchical residual-like way, cf. Figure 2. 3×3 convolution is followed by the input being split into s feature map subsets, indicated by the symbol X_i, where $i \in \{1, 2, ..., s\}$. Each feature subset X_i differs from the input feature map only in that it has $1/s$ fewer channels but the same spatial extent. With the exception of X_1, which is forwarded directly to the output, each X_i has a matching 3×3 convolution, indicated by $\mathrm{K}_i(\cdot)$. The output $\mathrm{K}_{i-1}(\cdot)$ from the earlier 3×3 convolution is then fed into $\mathrm{K}_i(\cdot)$ together with the feature subset X_i. The output of the module is produced by concatenating the outputs of all groups and forwarding them to a 1×1 convolution. Thus, Y_i can be:

$$Y_i = \begin{cases} X_i & i\text{=}1 \\ K_i(X_i) & i\text{=}2 \\ K_i(X_i + Y_{i-1}) & 2 < i \leq s \end{cases}$$

IV. Experimental Setup

We evaluate the performance of the proposed architecture on the ECAPA embedding on the development part of the VoxCeleb2 dataset with 5994 speakers as training data. VoxCeleb1 test set is taken into consideration as a validation set for hyperparameter optimization. As follow the baselines [5] [6], all models are trained using a standard Adam optimizer with cyclical learning rates ranging between 1e-8 and 1e-3. Using AAM-softmax with a margin of 0.2 and softmax prescaling of 30 for 4 cycles, all systems are trained.

A. Dataset

We use the development part of the VoxCeleb 2 [18] as our training set. This dataset contains over 1 million utterances for 5,994 speakers extracted from YouTube. The MUSAN [19] and RIR [20] datasets are used to generate extra samples for online data augmentation. VoxCeleb1 [21] has three types of evaluation trials, which are VoxCeleb1-O, VoxCeleb1-E and VoxCeleb1-H. For fairness of comparisons, we keep consistent with the ECAPA-TDNN and ECAPA CNN-TDNN experiments and choose VoxCeleb1-O as the validation set, this dataset contains 4,708 utterances from 40 speakers.

B. System Description

Both ECAPA-TDNN [5] and ECAPA CNN-TDNN [6] are used as baseline systems in this study. We can describe the proposed systems and the two baselines as follows:

- **ECAPA-TDNN (Re-implemented):** It follows the standard ECAPA-TDNN model from [5]. In the convolutional frame layers, there are 1024 channels, and the number of Res2Blocks is 3.

TABLE I
EER PERFORMANCE OF THE ECAPA-TDNN (ET) AND ECAPA CNN-TDNN (ECT) BASELINE MODELS AND PROPOSED ARCHITECTURES ON VOXCELEB1 TEST SET. PARAMETER s DEPICTS THE VALUE OF SCALE, g IS THE VALUE OF CARDINALITY, AND c IS THE NUMBER OF FILTERS.

Architecture	**Residual Units**	**Setting**	**No. Params**($Million$)	**EER**(%)	**PRI-ET**(%)	**PRI-ECT**(%)
ECAPA TDNN [5](Re-implemented)	Res2Net	8s×1024c	14.73	1.03		
ECAPA CNN-TDNN [6](Re-implemented)	ResNet	128c	27.54	0.97		
Extended ECAPA-TDNN	Res2Net	4s×1024c	15.43	1.12	-8.7	-15.5
		6s×1024c	14.96	1.07	-3.9	-10.3
	Res2NeXt	4s×4g × 1024c	14.17	1.02	+0.97	-5.2
		6s×8g × 1008c	**14.06**	**0.94**	**+8.7**	**+3.1**
		8s×8g × 1024c	13.87	1.03	0	-6.2
	ResNeXt	4g×1024c	16.00	1.12	-8.7	-15.5
		6g×1026c	15.23	1.13	-9.7	-16.5
		8g×1024c	14.87	1.29	-25.2	-32.99
RCT-Net	Improved ResNet	128c	27.54	0.98	+4.9	-1.03
	Res-TCN	128c	27.26	0.95	+7.8	+2.06
	Res2Net	4s×128c	27.03	0.98	+4.9	-1.03
		6s×128c	27.01	0.91	+11.7	+6.2
		8s×128c	27.01	0.94	+8.7	+3.1
	Res2NeXt	4s×4g × 128c	26.99	0.97	+5.8	0
		6s×8g × 144c	27.01	0.90	+12.6	+7.2
		8s×8g × 128c	**26.98**	**0.88**	**+14.6**	**+9.3**
	ResNeXt	4g×128c	27.12	1.11	-7.8	-14.4
		6g×132c	27.48	0.97	+5.8	0
		8g×128c	27.05	0.98	+4.9	-1.03

- **ECAPA CNN-TDNN (Re-implemented):** As proposed in [6] four layers of CNN are employed as a front-end for ECAPA-TDNN. Different from [6], we do not increase the intermediate channel dimension and depth in ECAPA-TDNN module, but the standard version with 3 SE-Res2Blocks and 1024 channels. This is for fair comparisons with ECAPA-TDNN and the proposed RCT-Net.
- **RCT-Net**: The standard ECAPA-TDNN with different residual blocks as a front-end.

C. Training the speaker embedding extractors

The input features are 80-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) extracted from a window length of 25 ms with a frame shift of 10 ms. Cepstral mean subtraction is used to normalize the two second random cropping of the MFCCs feature vectors. It is well known that data augmentation has great benefits for neural networks. So, we use the MUSAN (babble, music, noise, TV noise) corpora and the RIR corpora (reverb) for online data augmentation to generate five extra samples for each utterance. We apply SpecAugment [22] as the last step of augmentation, this algorithm randomly masks dimension of 10 and 8 in the temporal and frequency dimensions, respectively.

TABLE II
DIFFERENT SETTINGS OF *scale* AND *cardinality* DIMENSIONS ON MULTI-SCALE RESIDUAL BLOCKS

Residual Units	**Setting 1**	**Setting 2**	**Setting 3**
Res2Net	4s	6s	8s
ResNeXt	4g	6g	8g
Res2NeXt	4s×4g	6s×8g	8s×8g

V. RESULTS

A performance overview of the baseline systems described in Section II and our proposed architectures are summarized in Table I. We extend the baseline speaker embedding models by incorporating the proposed 1D and 2D convolutional stems with various bottleneck residual blocks. We then evaluate the Percent Relative Improvements (PRI) of the proposed models with the ECPA-TDNN and ECAPA CNN-TDNN baselines.

Results show that in general almost all RCT-based combinations (10 out of 11 combinations, i.e., around 91% of all combinations) lead to an improvement over standard ECAPA-TDNN. The results also demonstrate that all proposed models with potential to perform better than their corresponding baselines have fewer parameters. In the following, we analyze the performance in more detail wrt. to system combination constituents.

A. Variations in CNN stems representation

Further analyzing the results, we assume a competitive threshold of EER=1, i.e., a high-performance system threshold where the amount of falsely rejected and falsely accepted speakers in an ASV system would be equally high, namely 1%. Accordingly, as shown in Table I, while 87.5% of any ECAPA-TDNN extension included in the experiments are above the threshold of 1%, 91% of RCT-Net proposed models are below it. We could therefore assume that overall the 2D convolutional stems are more optimally suited for the representation of speaker embeddings for ASV systems, compared to 1D representations.

B. Dimension variations

Findings of prior benchmark experiments [4] imply that scale is an effective dimension to enhance model performance.

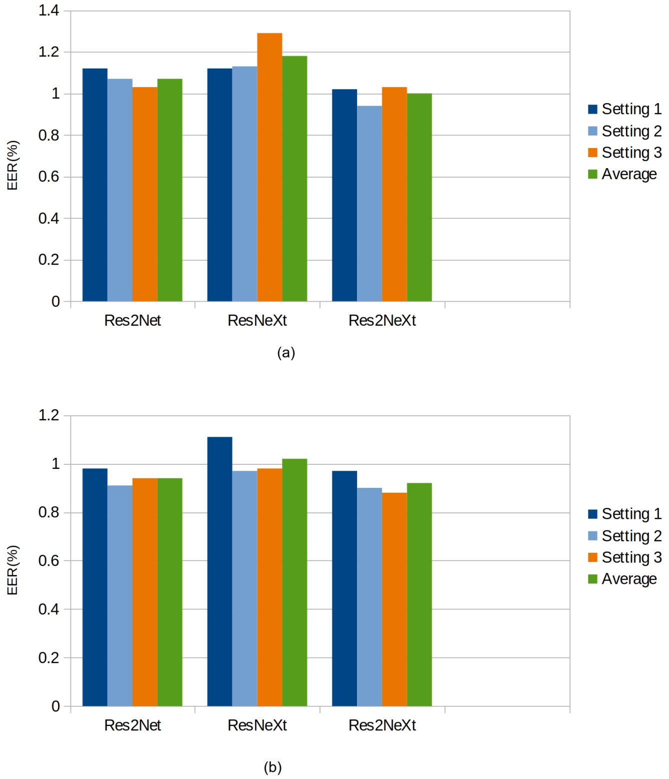
Moreover, scaling up is more efficient than other dimensions. In general, this finding can be confirmed, as for most system configurations s=4 results in inferior performance, compared to higher values. However, rising the scale from 6 to 8 does not always lead to gain. On this level, the overall performance also depends on the remaining parameters c and g.

C. Multi-scale residual blocks

In terms of EER, the best model using Res2NeXt$-8s \times 8g \times 128c$ surpasses both ECAPA-TDNN and ECAPA CNN-TDNN baselines by 14.6% and 8.7%, respectively. Remarkably, Res2NeXt$-6s \times 8g \times 1008c$ even outperforms the baseline, ResNet-128c, with only 51% of the number of parameters in the model (see Table I). As shown in Figure 3, for 1D representations the introduction of multi-scale blocks in ResNeXt alone does not lead to any improvement. However, when combining the advantages of it into the Res2NeXt model, the performance significantly improves, i.e., by 8.7% - a performance value even outperforming the ECAPA CNN-TDNN baseline operating on a 2D representation in the stem. For the RCT-Net based models, the introduction of multi-scale blocks clearly improves the overall performance, with only the exception of ResNeXt model with too small scale settings discussed above. All models show significant improvement, best of which improves performance by 14.6% using a Res2NeXt block. Eventually, we can hypothesize that the multi-scale feature setup greatly benefits from the 2D convolution processing in the entrance of the stem.

VI. Discussions

Based on our results, we can conclude that integrating 2D Res2NeXt with TDNN is the best combination of two strong structures of TDNN and residual blocks. As a result, in our experiments representing features at multiple scales and constructing hierarchical residual-like connections within a single residual block in dimensions of both scale and cardinality is more performant than without or standalone dimensions of either scale or cardinality. A possible explanation could be the difference in the approach to obtaining multi-scale features in different residual-based architectures. Res2Net, for example, splits the original input into multiple groups according to the channels. The output of one group is fed into the next group, and so on, and all segments are concatenated as the final result. On the other side, Res2NeXt, repeats a building block that aggregates a set of transformations with the same topology and expands the range of receptive fields for each network layer, and depicts multi-scale features at a finer level. Accordingly, by integrating hierarchical multi-scale feature representation within the bottleneck module, the multi-scale feature representation is improved at both the global and local levels. Finally, in our experiment, the joint benefits of a parallel stacking layer of ResNeXt rather than sequential layers of standard ResNet architectures, multi-scaling features in Res2Net, and expanding the range of receptive fields show the potential to extract more invariant feature representations in a joint Res2NeXt architecture.



(a)

(b)

Fig. 3. Impact of various *scale* and *cardinality* dimensions with different settings as indicated in Table II. (a) ECAPA-TDNN based experiments, (b) ECAPA CNN-TDNN based experiments.

VII. Conclusion

In this study, we adapt the frame-level layer architecture that integrates multiple ideas motivated by the convolutional block and multi-scale architectures. In our experiments, we evaluate the performance of integrating different residual blocks into TDNN-based structures. The best model using Res2NeXt improves current state-of-the-art by 14.6% relative on VoxCeleb1 test set.

These promising findings motivate us to investigate hybrid architectures in more detail and propose structures to reduce computational complexity in our upcoming studies. We will continue to evaluate the performance of various residual unit types as we integrate them with the 2D ECAPA-TDNN representation and explore several directions of multimodal fusion approaches. We will also provide speech-level interpretation of the proposed TDNN-based architectures for understanding our models. This includes visualizing the acoustic concepts the model has learned and comparing how they are represented in the model layers using [23] [24], etc, and generalizing our findings with more data utilizing additional datasets and evaluation metrics such as Minimum Value of Detection Cost Function (MinDCF).

Acknowledgement

[This research has been partly funded by the Federal Ministry of Education and Research of Germany in the project

Emonymous (project number S21060A) and partly funded by the Volkswagen Foundation in the project AnonymPrevent (AI-based Improvement of Anonymity for Remote Assessment, Treatment and Prevention against Child Sexual Abuse).

References

[1] 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.

[2] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," CoRR, vol. abs/1602.07261, 2016. [Online]. Available: http://arxiv.org/abs/1602.07261.

[3] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492–1500.

[4] S.-H. Gao, et al. "Res2net: A new multi-scale backbone architecture," IEEE transactions on pattern analysis and machine intelligence, vol. 43, no. 2, pp. 652–662, 2019.

[5] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA- TDNN: Emphasized Channel Attention, Propagation and Aggrega- tion in TDNN Based Speaker Verification," in Proc. Interspeech 2020, 2020, pp. 3830–3834.

[6] J. Thienpondt, B. Desplanques, and K. Demuynck, "Integrating frequency translational invariance in TDNNs and frequency positional information in 2d ResNets to enhance speaker verification," in Interspeech 2021. ISCA, aug 2021. [Online]. Available: $https : //doi.org/10.21437\%2Finterspeech.2021 - 1570$

[7] T. Liu, R. K. Das, K. A. Lee, and H. Li, "MFA: TDNN with multi-scale frequency-channel attention for textindependent speaker verification with short utterances," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022.

[8] Zhang, Y.-J., et al. (2021) Improving Time Delay Neural Network Based Speaker Recognition with Convolutional Block and Feature Aggregation Methods. Proc. Interspeech 2021, 76-80, doi: 10.21437/Interspeech.2021-356.

[9] H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin, "Exploiting deep residual networks for human action recognition from skeletal data," CoRR, vol. abs/1803.07781, 2018. [Online]. Available: http://arxiv.org/abs/1803.07781

[10] R. Khamsehashari, K. Gadzicki, and C. Zetzsche, "Deep residual temporal convolutional networks for skeleton-based human action recognition," in Computer Vision Systems, D. Tzovaras, D. Gi- akoumis, M. Vincze, and A. Argyros, Eds. Cham: Springer International Publishing, 2019, pp. 376–385.

[11] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in Proc. ICASSP, 2018, pp. 5329–5333.

[12] 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.

[13] R. Khamsehashari, et al. "Voice Privacy - leveraging multi-scale blocks with ECAPA-TDNN SE-Res2NeXt extension for speaker anonymization," in Proc. 2nd Symposium on Security and Privacy in Speech Communication, 2022, pp. 43–48.

[14] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," CoRR, vol. abs/1611.05267, 2016. [Online]. Available: http://arxiv.org/abs/1611.05267

[15] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," CoRR, vol. abs/1704.04516, 2017. [Online]. Available: http://arxiv.org/abs/ 1704.04516

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014. [Online]. Available: https://arxiv.org/abs/1409.1556

[17] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[18] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," arXiv preprint arXiv:1806.05622, 2018.

[19] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.

[20] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 5220–5224.

[21] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," arXiv preprint arXiv:1706.08612, 2017.

[22] D. S. Park, et al. "Specaugment: A simple data augmen- tation method for automatic speech recognition," arXiv preprint arXiv:1904.08779, 2019.

[23] O. Ozyegen, I. Ilic, and M. Cevik, Evaluation of interpretability methods for multivariate time series forecasting. Appl Intell 52, 4727–4743 (2022). https://doi.org/10.1007/s10489-021-02662-2

[24] R. R. Selvaraju, et al. "Grad-CAM: visual explanations from deep networks via gradient-based localization," in IEEE international conference on computer vision, 2017, pp. 618–626.