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

Razieh Khamsehashari[1], Fengying Miao[1], Tim Polzehl[2], Sebastian Möller[1]

Quality and Usability Lab, TU Berlin, Germany[1]

Speech and Language Technology Lab, DFKI Berlin, Germany[2]

Email: razieh.khamsehashari@tu-berlin.de

Academic Researcher

Technical University of Berlin

Quality and Usability Lab Institute of Software Engineering and Theoretical Computer Science

Faculty IV Electrical Engineering and Computer Science

Ernst-Reuter-Platz 7

TU-Hochhaus, Sekr. TEL 18

10587 Berlin, Germany

Email: razieh.khamsehashari@tu-berlin.de

Biography

Razieh Khamsehahsari received her master’s degree in Artificial Intelligence from Isfahan University of Technology with a thesis on “Object Retrieval and Recognition of Digital Images Based on Brain Computer Interface and Computer Vision”. She has also a Bachelor's degree in Software Engineering. Studying in a multidisciplinary environment has provided her with a strong background in various areas, including BCI, cognitive computational neuroscience, speech processing and computer vision. Razieh has worked as a research assistant in the Quality and Usability Lab, Technical University of Berlin since April 2022 while pursuing her Ph.D. She is working on deep learning based speech processing applications.

Interested Research Fields

- Deep Learning for Speech and Language Processing Applications
- Multisensorial Interaction at Machine Perception

Introduction

- 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 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.

Methods

- Two types of TDNN-based speaker embedding models, ECAPA-TDNN [1] and ECAPA CNN-TDNN [2], are considered as reliable baselines to evaluate the performance of our suggested architectures.

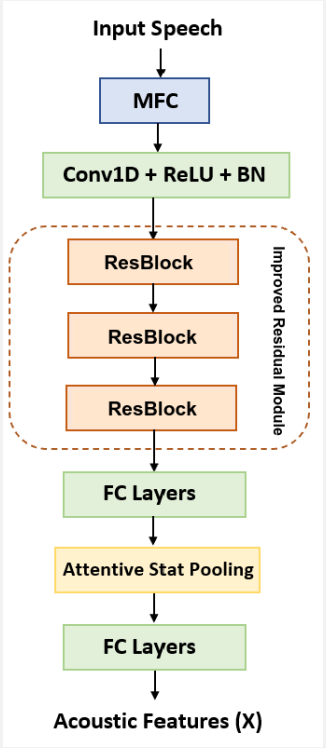
- Experiment with the convolutional stem, including various bottleneck residual blocks such as Res2Net [3], Res2NeXt [4], standard ResNet [5], Improved ResNet [6] and ResTCN [7].

- Our proposed architectures:
  - Extended ECAPA-TDNN
  - RCT-Net

System Architectures

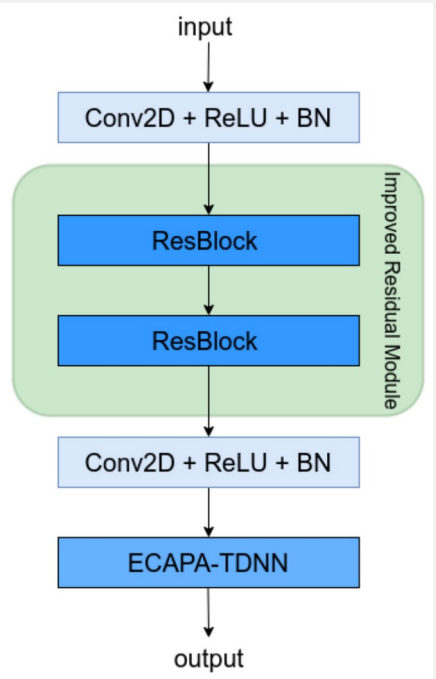
Extended ECAPA-TDNN

- Evaluate the effectiveness of the baseline model using Various CNN dimensions including scale and cardinality

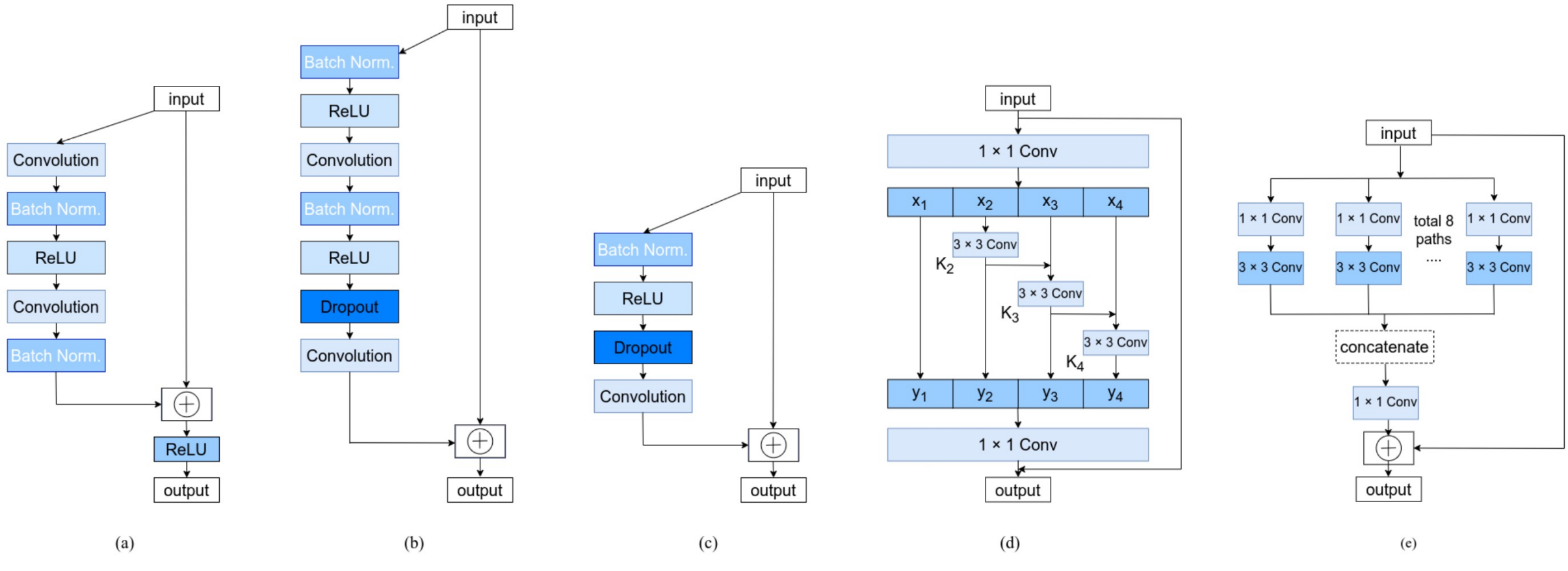


RCT-Net

- 2D convolutional stem for the ECAPA-TDNN speaker verification model
- Incorporating frequency translational invariance in the initial layers of the network



Residual blocks



The structures of bottleneck residual blocks in different architectures. Standard residual blocks in (a) ResNet [5], (b) Improved ResNet [6], and (c) Res-TCN [8]. Multi-scale residual blocks in (d) Res2Net [3] and (e) ResNeXt. [4]

Experimental Setup

Dataset

- Evaluate on development part of VoxCeleb2 dataset with 5994 speakers as training data.
- MUSAN and RIR datasets to generate extra samples for data augmentation.
- VoxCeleb1-O test set contains 4,708 utterances from 40 speakers as validation set.

Training

- The input features are 80-dimensional MFCCs extracted from a window length of 25 ms with a frame shift of 10 ms.
- Standard Adam optimizer with cyclical learning rates ranging between 1e-8 and 1e-3.
- AAM-softmax with a margin of 0.2 and softmax prescaling of 30 for 4 cycles.

Results

- Almost all RCT-based combinations (~91%) lead to an improvement over standard ECAPA-TDNN.
- All proposed models with potential to perform better than their corresponding baselines have fewer parameters.
- The best model using Res2NeXt−8s ×8g × 128c surpasses both ECAPA-TDNN and ECAPA CNN-TDNN baselines by **14.6%** and **8.7%**, respectively. Remarkably, Res2NeXt−6s × 8g × 1008c even outperforms the baseline, ResNet-128c, with only **51%** of the number of parameters in the model.

Architecture	**Residual Units**	**Setting**	**No. Params**$(Million)$	**EER**$(\%)$	**PRI-ET**$(\%)$	**PRI-ECT**$(\%)$
ECAPA TDNN [5](Re-implemented)	Res2Net	$8s\times1024c$	14.73	1.03		
ECAPA CNN-TDNN [6](Re-implemented)	ResNet	128c	27.54	0.97		
Extended ECAPA-TDNN	Res2Net	$4s\times1024c$	15.43	1.12	-8.7	-15.5
		$6s\times1024c$	14.96	1.07	-3.9	-10.3
	Res2NeXt	$4s\times4g\times1024c$	14.17	1.02	+0.97	-5.2
		$6s\times8g\times1008c$	**14.06**	**0.94**	**+8.7**	**+3.1**
		$8s\times8g\times1024c$	13.87	1.03	0	-6.2
	ResNeXt	$4g\times1024c$	16.00	1.12	-8.7	-15.5
		$6g\times1026c$	15.23	1.13	-9.7	-16.5
		$8g\times1024c$	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	$4s\times128c$	27.03	0.98	+4.9	-1.03
		$6s\times128c$	27.01	0.91	+11.7	+6.2
		$8s\times128c$	27.01	0.94	+8.7	+3.1
	Res2NeXt	$4s\times4g\times128c$	26.99	0.97	+5.8	0
		$6s\times8g\times144c$	27.01	0.90	+12.6	+7.2
		$8s\times8g\times128c$	**26.98**	**0.88**	**+14.6**	**+9.3**
	ResNeXt	$4g\times128c$	27.12	1.11	-7.8	-14.4
		$6g\times132c$	27.48	0.97	+5.8	0
		$8g\times128c$	27.05	0.98	+4.9	-1.03

Results

Variations in CNN stems representation:

- 2D convolutional stems are more optimally suited for the representation of speaker embedding compared to 1D representations.
- 87.5% of any ECAPA-TDNN extension included in the experiments are above the threshold of 1%, 91% of RCT-Net models are below it.

Architecture	**Residual Units**	**Setting**	**No. Params**$(Million)$	**EER**$(\%)$	**PRI-ET**$(\%)$	**PRI-ECT**$(\%)$
ECAPA TDNN [5](Re-implemented)	Res2Net	$8s\times1024c$	14.73	1.03		
ECAPA CNN-TDNN [6](Re-implemented)	ResNet	128c	27.54	0.97		
Extended ECAPA-TDNN	Res2Net	$4s\times1024c$	15.43	1.12	-8.7	-15.5
		$6s\times1024c$	14.96	1.07	-3.9	-10.3
	Res2NeXt	$4s\times4g\times1024c$	14.17	1.02	+0.97	-5.2
		$6s\times8g\times1008c$	**14.06**	0.94	**+8.7**	**+3.1**
		$8s\times8g\times1024c$	13.87	1.03	0	-6.2
	ResNeXt	$4g\times1024c$	16.00	1.12	-8.7	-15.5
		$6g\times1026c$	15.23	1.13	-9.7	-16.5
		$8g\times1024c$	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	$4s\times128c$	27.03	0.98	+4.9	-1.03
		$6s\times128c$	27.01	0.91	+11.7	+6.2
		$8s\times128c$	27.01	0.94	+8.7	+3.1
	Res2NeXt	$4s\times4g\times128c$	26.99	0.97	+5.8	0
		$6s\times8g\times144c$	27.01	0.90	+12.6	+7.2
		$8s\times8g\times128c$	**26.98**	0.88	**+14.6**	**+9.3**
	ResNeXt	$4g\times128c$	27.12	1.11	-7.8	-14.4
		$6g\times132c$	27.48	0.97	+5.8	0
		$8g\times128c$	27.05	0.98	+4.9	-1.03

Results

Dimension variations:

- Findings of prior benchmark experiments [3] imply that scaling up is more efficient than other dimensions.
- This finding can be confirmed, as for most system configurations *s*=4 results in inferior performance, compared to higher values.
- On this level, the overall performance also depends on the remaining parameters *c* and *g*.

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

Results

Dimension variations:

- Findings of prior benchmark experiments [3] imply that scaling up is more efficient than other dimensions.
- This finding can be confirmed, as for most system configurations *s*=4 results in inferior performance, compared to higher values.
- On this level, the overall performance also depends on the remaining parameters *c* and *g*.

Architecture	**Residual Units**	**Setting**	**No. Params**$(Million)$	**EER**$(\%)$	**PRI-ET**$(\%)$	**PRI-ECT**$(\%)$
ECAPA TDNN [5](Re-implemented)	Res2Net	$8s\times1024c$	14.73	1.03		
ECAPA CNN-TDNN [6](Re-implemented)	ResNet	128c	27.54	0.97		
Extended ECAPA-TDNN	Res2Net	$4s\times1024c$	15.43	1.12	-8.7	-15.5
		$6s\times1024c$	14.96	1.07	-3.9	-10.3
	Res2NeXt	$4s\times4g\times1024c$	14.17	1.02	+0.97	-5.2
		$6s\times8g\times1008c$	**14.06**	**0.94**	**+8.7**	**+3.1**
		$8s\times8g\times1024c$	13.87	1.03	0	-6.2
	ResNeXt	$4g\times1024c$	16.00	1.12	-8.7	-15.5
		$6g\times1026c$	15.23	1.13	-9.7	-16.5
		$8g\times1024c$	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	$4s\times128c$	27.03	0.98	+4.9	-1.03
		$6s\times128c$	27.01	0.91	+11.7	+6.2
		$8s\times128c$	27.01	0.94	+8.7	+3.1
	Res2NeXt	$4s\times4g\times128c$	26.99	0.97	+5.8	0
		$6s\times8g\times144c$	27.01	0.90	+12.6	+7.2
		$8s\times8g\times128c$	**26.98**	**0.88**	**+14.6**	**+9.3**
	ResNeXt	$4g\times128c$	27.12	1.11	-7.8	-14.4
		$6g\times132c$	27.48	0.97	+5.8	0
		$8g\times128c$	27.05	0.98	+4.9	-1.03

Results

Dimension variations:

- Findings of prior benchmark experiments [3] imply that scaling up is more efficient than other dimensions.
- This finding can be confirmed, as for most system configurations *s*=4 results in inferior performance, compared to higher values.
- On this level, the overall performance also depends on the remaining parameters *c* and *g*.

Architecture	**Residual Units**	**Setting**	**No. Params**$(Million)$	**EER**$(\%)$	**PRI-ET**$(\%)$	**PRI-ECT**$(\%)$
ECAPA TDNN [5](Re-implemented)	Res2Net	$8s\times1024c$	14.73	1.03		
ECAPA CNN-TDNN [6](Re-implemented)	ResNet	128c	27.54	0.97		
Extended ECAPA-TDNN	Res2Net	$4s\times1024c$	15.43	1.12	-8.7	-15.5
		$6s\times1024c$	14.96	1.07	-3.9	-10.3
	Res2NeXt	$4s\times4g\times1024c$	14.17	1.02	+0.97	-5.2
		$6s\times8g\times1008c$	**14.06**	**0.94**	**+8.7**	**+3.1**
		$8s\times8g\times1024c$	13.87	1.03	0	-6.2
	ResNeXt	$4g\times1024c$	16.00	1.12	-8.7	-15.5
		$6g\times1026c$	15.23	1.13	-9.7	-16.5
		$8g\times1024c$	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	$4s\times128c$	27.03	0.98	+4.9	-1.03
		$6s\times128c$	27.01	0.91	+11.7	+6.2
		$8s\times128c$	27.01	0.94	+8.7	+3.1
	Res2NeXt	$4s\times4g\times128c$	26.99	0.97	+5.8	0
		$6s\times8g\times144c$	27.01	0.90	+12.6	+7.2
		$8s\times8g\times128c$	**26.98**	**0.88**	**+14.6**	**+9.3**
	ResNeXt	$4g\times128c$	27.12	1.11	-7.8	-14.4
		$6g\times132c$	27.48	0.97	+5.8	0
		$8g\times128c$	27.05	0.98	+4.9	-1.03

Results

Dimension variations:

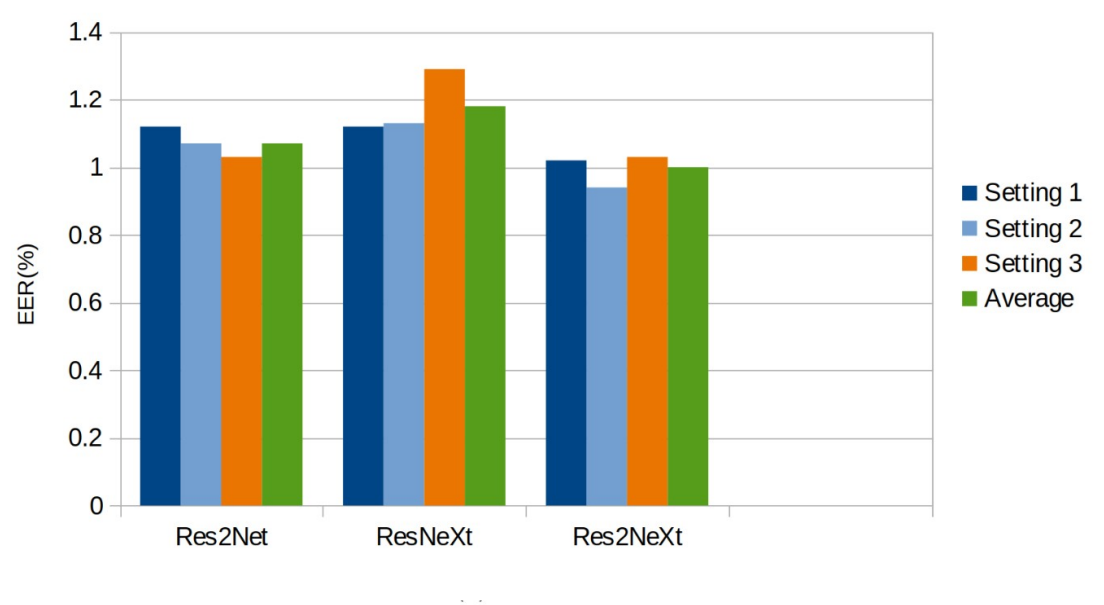
- Findings of prior benchmark experiments [3] imply that scaling up is more efficient than other dimensions.
- This finding can be confirmed, as for most system configurations *s*=4 results in inferior performance, compared to higher values.
- On this level, the overall performance also depends on the remaining parameters *c* and *g*.

Architecture	**Residual Units**	**Setting**	**No. Params**$(Million)$	**EER**$(\%)$	**PRI-ET**$(\%)$	**PRI-ECT**$(\%)$
ECAPA TDNN [5](Re-implemented)	Res2Net	$8s\times1024c$	14.73	1.03		
ECAPA CNN-TDNN [6](Re-implemented)	ResNet	128c	27.54	0.97		
Extended ECAPA-TDNN	Res2Net	$4s\times1024c$	15.43	1.12	-8.7	-15.5
		$6s\times1024c$	14.96	1.07	-3.9	-10.3
	Res2NeXt	$4s\times4g\times1024c$	14.17	1.02	+0.97	-5.2
		$6s\times8g\times1008c$	**14.06**	**0.94**	**+8.7**	**+3.1**
		$8s\times8g\times1024c$	13.87	1.03	0	-6.2
	ResNeXt	$4g\times1024c$	16.00	1.12	-8.7	-15.5
		$6g\times1026c$	15.23	1.13	-9.7	-16.5
		$8g\times1024c$	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	$4s\times128c$	27.03	0.98	+4.9	-1.03
		$6s\times128c$	27.01	0.91	+11.7	+6.2
		$8s\times128c$	27.01	0.94	+8.7	+3.1
	Res2NeXt	$4s\times4g\times128c$	26.99	0.97	+5.8	0
		$6s\times8g\times144c$	27.01	0.90	+12.6	+7.2
		$8s\times8g\times128c$	**26.98**	**0.88**	**+14.6**	**+9.3**
	ResNeXt	$4g\times128c$	27.12	1.11	-7.8	-14.4
		$6g\times132c$	27.48	0.97	+5.8	0
		$8g\times128c$	27.05	0.98	+4.9	-1.03

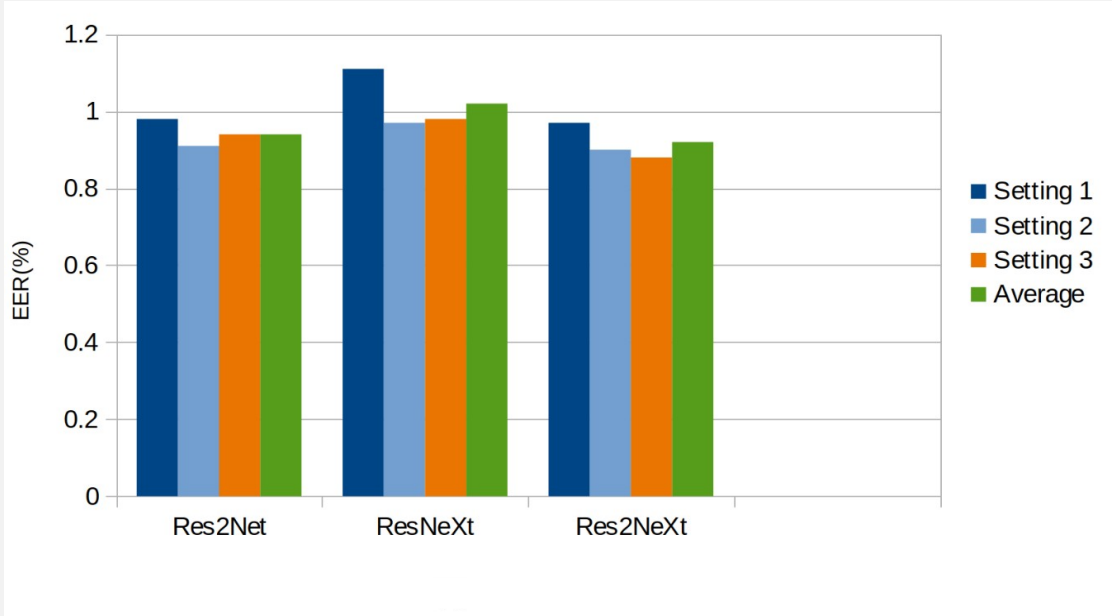
Results

Multi-scale residual blocks:

ECAPA-TDNN based experiments



ECAPA CNN-TDNN based experiments



Residual Units	Setting 1	Setting 2	Setting 3
Res2Net	4s	6s	8s
ResNeXt	4g	6g	8g
Res2NeXt	$4s \times 4g$	$6s \times 8g$	$8s \times 8g$

Results

ECAPA-TDNN based experiments:

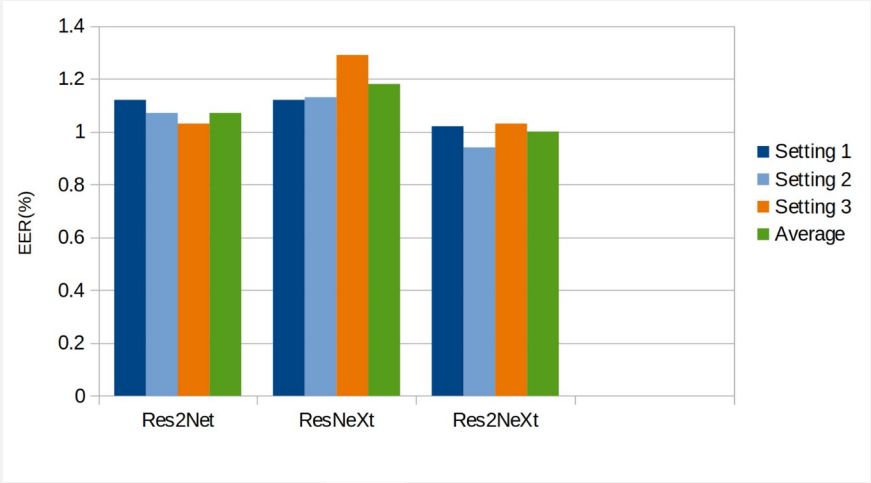
- For 1D representations the introduction of multi-scale blocks in ResNeXt alone does not lead to any improvement.
- When combining it into the Res2NeXt model, the performance improves by 8.7%.

ECAPA CNN-TDNN based experiments:

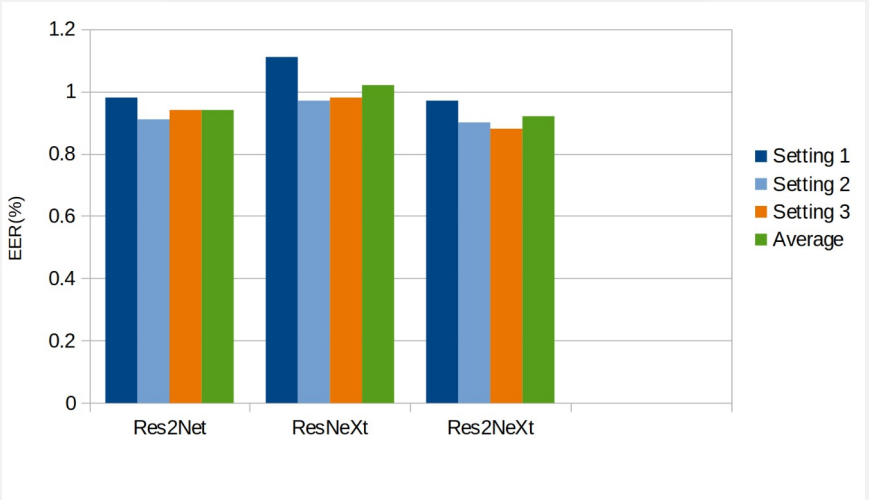
- Introduction of multi-scale blocks clearly improves the overall performance.
- We can hypothesize that the multi-scale feature setup greatly benefits from the 2D convolution processing in the entrance of the stem.

Residual Units	Setting 1	Setting 2	Setting 3
Res2Net	4s	6s	8s
ResNeXt	4g	6g	8g
Res2NeXt	$4s \times 4g$	$6s \times 8g$	$8s \times 8g$

ECAPA-TDNN based experiments



ECAPA CNN-TDNN based experiments



Discussions

- Based on our results, integrating 2D Res2NeXt with TDNN is the best combination of two strong structures of TDNN and residual blocks.
- 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.

Conclusion

- This study adapt the frame-level layer architecture that integrates multiple ideas motivated by the convolutional block and multi-scale architectures.
- The best model using Res2NeXt improves current state-of-the-art by 14.6% relative on VoxCeleb1 test set.

Future Works

- Investigate hybrid architectures in more details and propose structures to reduce computational complexity
- Speech-level interpretation of the proposed TDNN-based architectures
  - Visualizing the acoustic concepts using Explainable AI methods
  - Generalizing our findings with additional datasets and evaluation metrics

References

[1] 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.

[2] 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%2F interspeech.2021 − 1570.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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

THANK YOU

For Your Attention