Voice Privacy - Leveraging Multi-Scale Blocks with ECAPA-TDNN SE-Res2NeXt Extension for Speaker Anonymization

Razieh Khamsehashari, Yamini Sinha, Jan Hintz, Suhita Ghosh, Tim Polzehl, Carlos Franzreb, Sebastian Stober, Ingo Siegert

Quality and Usability Lab, TU Berlin, Germany

Speech and Language Technology Lab, DFKI Berlin, Germany

Mobile Dialog Systems, Otto von Guericke University Magdeburg, Germany

Artificial Intelligence Lab (AILab), Otto von Guericke University Magdeburg, Germany

Prevention Network Dunkenfeld [3]

“You are not guilty because of your sexual desire, but you are responsible for your sexual behavior. There is help! Don’t become an offender!”

AnonymPrevent

- Video-calls anonymized in real-time.
- Callers are protected from their countries.
- The system may lead to more acceptance.

Own Research

Main ideas

- Add more low-level descriptors
- Apply other speech representations
- Extend x-vectors, e.g. ECAPA
- Use HiFI-GAN and StarGAN for conversion
- Fuse conversion and NSF

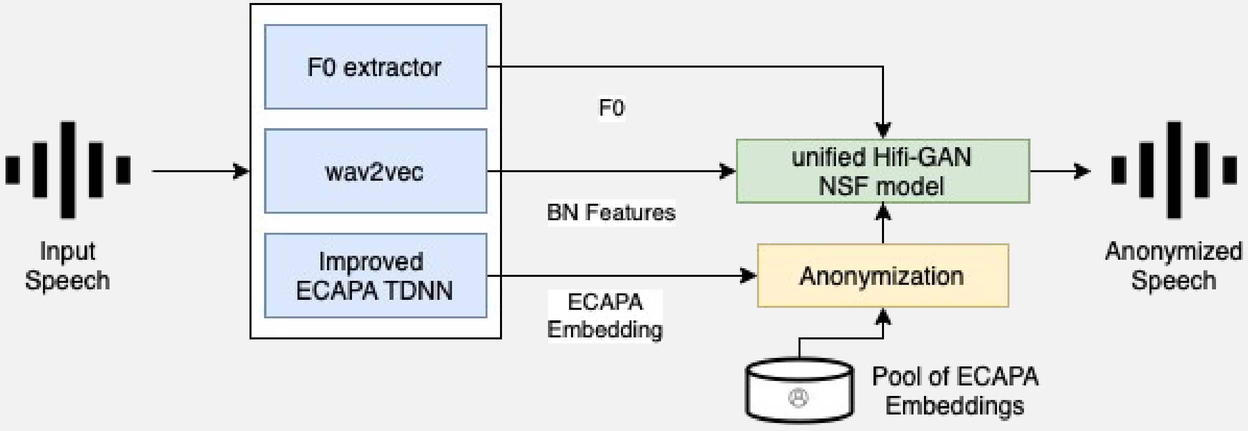


Figure 1: Adaptation of Baseline

Experiments in the SPSC Paper

- ECAPA vs. x-vectors
- Improved ECAPA-TDNN (SE-Res2NeXt)
- Extend 1d conv to 2d conv as input to ECAPA

Speaker Model: ECAPA-TDNN

- Enhanced version of the x-vector topology
- Introduces 1-dimensional TDNN-specific SE-blocks

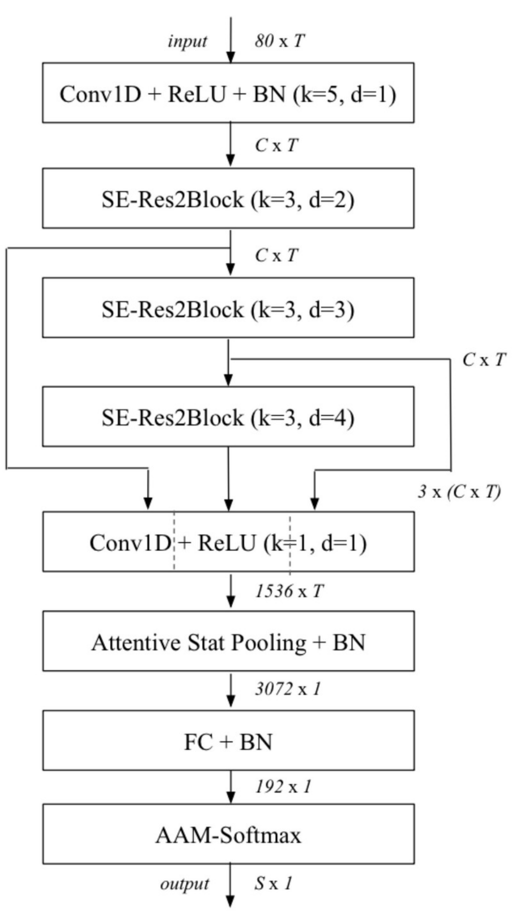


Figure 2: Network topology of the ECAPA-TDNN [1]

Extended ECAPA-TDNN

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

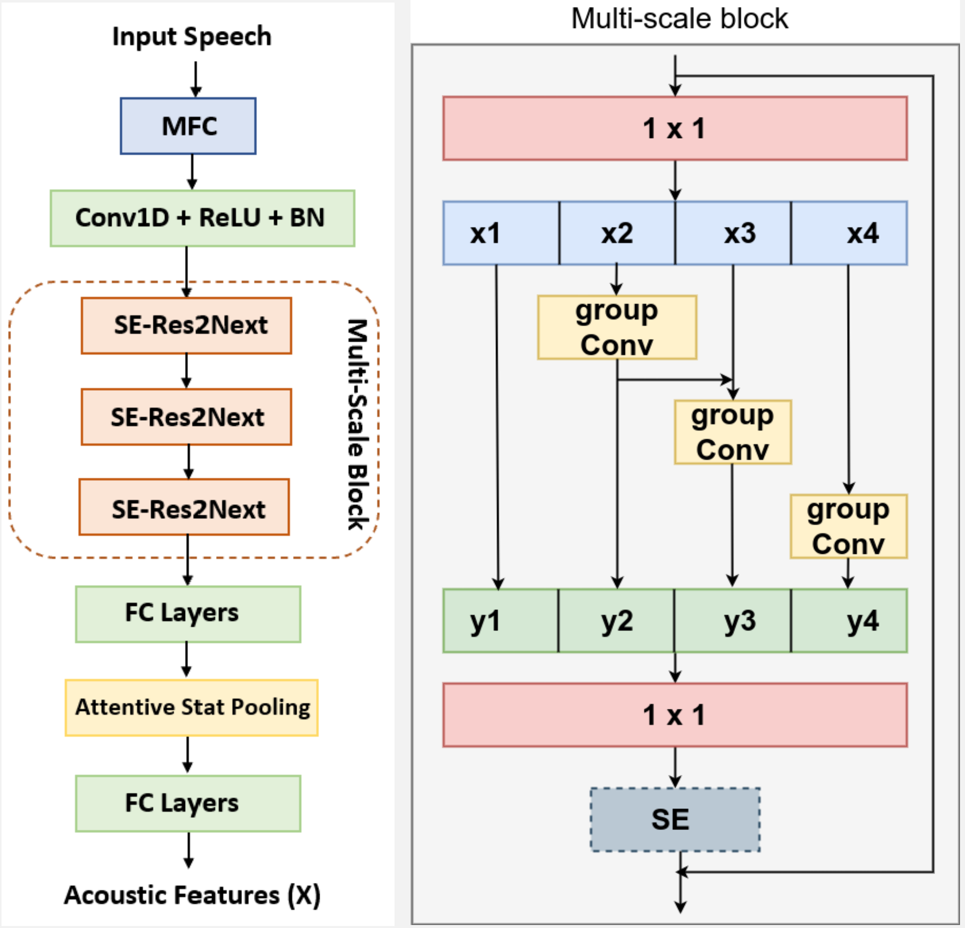


Figure 3: Extended ECAPA (SE-Res2NeXt architecture)

Training ECAPA Embedding Extractor

Dataset

- Evaluate on the development part of the VoxCeleb2 dataset with 5994 speakers as training data
- Use VoxCeleb1 test set as a validation set

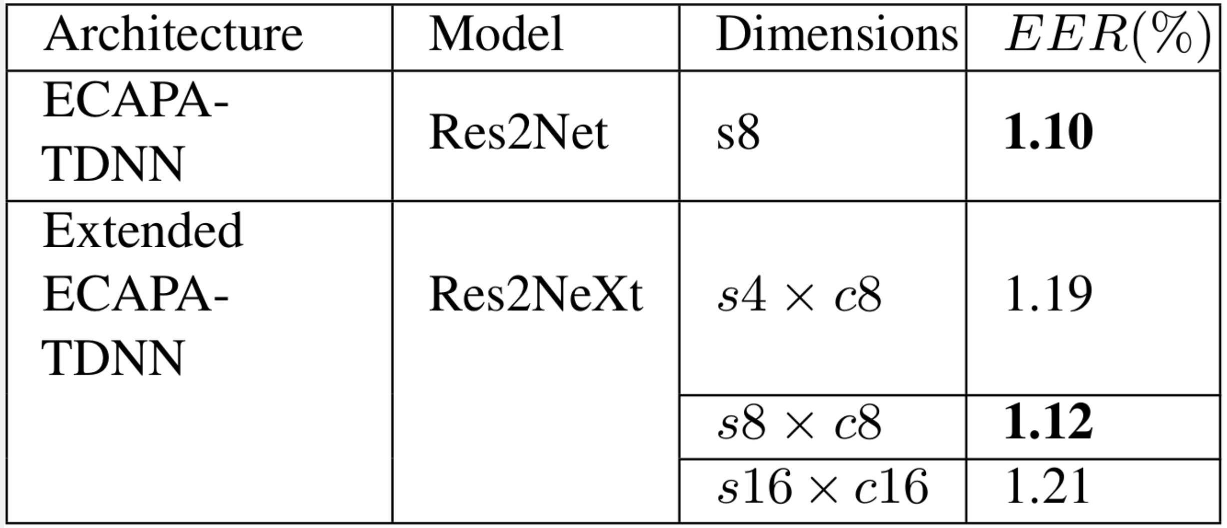
Training

- 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

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

Table 1: Top-1 test error (%) for the VoxCeleb dataset.



Architecture	Model	Dimensions	$EER(\%)$
ECAPA-TDNN	Res2Net	s8	**1.10**
Extended ECAPA-TDNN	Res2NeXt	$s4 \times c8$	1.19
		$s8 \times c8$	**1.12**
		$s16 \times c16$	1.21

ECAPA CNN-TDNN

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

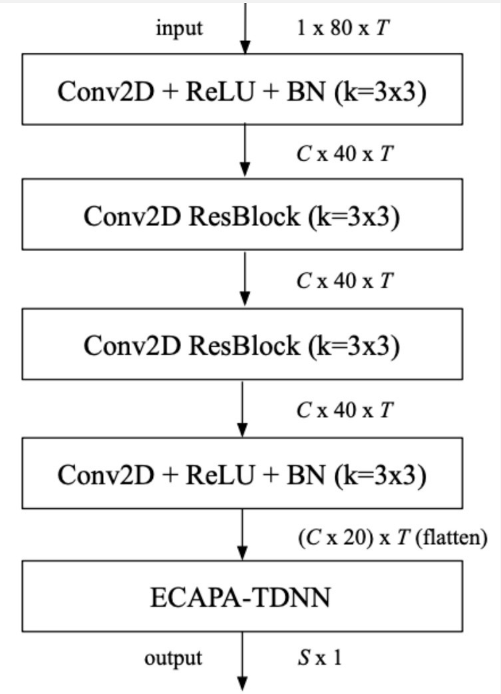
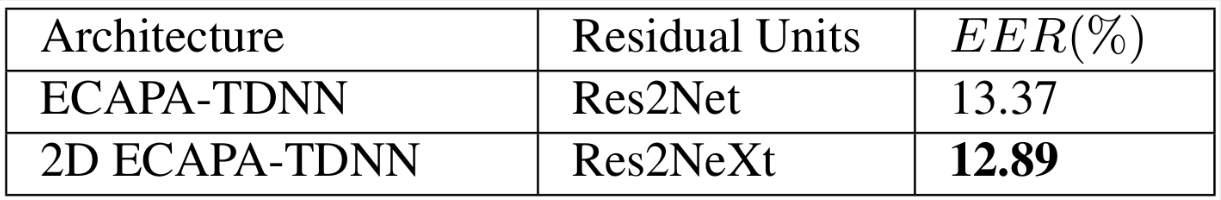


Figure 4: The 2D convolutional stem of the ECAPA CNN-TDNN architecture [2].

Preliminary Results

- Evaluate with small subset of the development part of the VoxCeleb2 dataset
- Using a 2D ECAPA-TDNN with Res2NeXt residual units improves the preliminary EER results by roughly 0.5% absolute

Table 2: Top-1 test error (%) for the VoxCeleb dataset.



Architecture	Residual Units	$EER(\%)$
ECAPA-TDNN	Res2Net	13.37
2D ECAPA-TDNN	Res2NeXt	**12.89**

Conclusion

- This study presents an extended ECAPA-TDNN and 2D ECAPA-TDNN with Res2XBlock integration for speaker verification
- Extending ECAPA-TDNN with 1-dimensional TDNN-specific SE-blocks does not improve by adding an extra dimension of cardinality
- changing to 2D ECAPA-TDNN we reach a relative improvement of roughly 0.5% absolute in EER over a strong baseline

Future Works

- We will keep evaluating the effectiveness of different types of residual units while integrating them with the 2D ECAPA-TDNN representation
- Using more data utilizing additional datasets and generating extra samples for each utterance by data augmentation

References

[1] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation 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%2Finterspeech.2021-1570

[3] K. Beier, “The German Dunkelfeld Project:Proactive Strategies to Prevent Child Sexual Abuse and the Use of Child Abusive Images” Contributions from the 8th Annual International Forum 2014

THANK YOU

For Your Attention