Informed Target Speaker Extraction Using TCN and TCN-Conformer Architectures for the 2nd Clarity Enhancement Challenge

Marvin Tammen[1], *Ragini Sinha*[2], *Henri Gode*[1], *Daniel Fejgin*[1], *Wiebke Middelberg*[1], *Eike J. Nustede*[1], *Reza Varzandeh*[1], *Jörn Anemüller*[1], *Gerald Enzner*[1], *Simon Doclo*[1,2]

[1]Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all,
University of Oldenburg, Germany
[2]Fraunhofer Institute for Digital Media Technology IDMT,
Oldenburg Branch for Hearing, Speech and Audio Technology HSA, Germany

Abstract

In this contribution, we present two deep neural network (DNN)-based systems submitted to the 2nd Clarity Enhancement Challenge, aiming at improving speech intelligibility of the target speaker for hearing-impaired listeners in a reverberant acoustic scenario with a target speaker and multiple interfering sources. The systems combine a DNN-based speaker-informed target speaker extraction stage to estimate the target speaker from the mixture and an audiogram-based hearing loss compensation stage. An objective evaluation based on the hearing aid speech perception index (HASPI) metric shows that both submitted systems result in a significant improvement in terms of speech intelligibility compared with the mixture signals as well as the baseline system.

1. Introduction

In the 2nd Clarity Enhancement Challenge (CEC2), a hearing-impaired listener wearing binaural hearing aids equipped with three microphones each is considered in a mildly reverberant room with a target speaker and two or three interfering sources. The listener is initially oriented away from the target speaker, but turns his head towards the target speaker once he/she starts talking. We propose to tackle this challenging scenario with two systems, each consisting of two cascaded blocks, i.e., a DNN-based speaker-informed target speaker extraction stage and a hearing loss compensation stage. In the following subsections, we will first provide some general background information about these two blocks, before presenting the details of the proposed systems in Section 2.

1.1. Speaker Informed Target Speaker Extraction

A speaker-informed target speaker extraction system typically consists of a speaker embedder network and a speaker separator network [1]. The speaker embedder network generates an embedding from the reference speech of the target speaker, which is an utterance of the target speaker different from the utterance of the target speaker in the mixture. The speaker separator network aims at estimating the target speaker from the mixture guided by the embedding generated using the speaker embedder network. The speaker embedder and separator networks can be trained either separately [2] or jointly [3].

1.2. Hearing Loss Compensation

To compensate for potential hearing loss of the listener, we apply the CEC2 baseline system as provided by the challenge organizers.

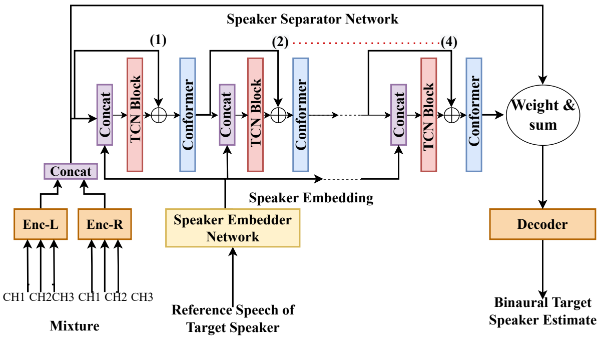

Figure 1: *Proposed system CEC_E036 based on TCN-Conformers.*

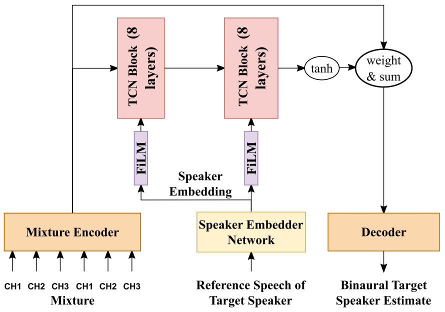

Figure 2: *Proposed system CEC_E038 based on TCNs.*

2. Proposed Systems

In this section, we describe our proposed speaker-informed target speaker extraction systems:

- **CEC2_E036**: TCN-Conformer-based system.
- **CEC2_E038**: TCN-based system.

Both systems process the signals in the time domain, where the speaker embedder and separator networks are trained jointly. The speaker separator network consists of three blocks: mixture encoder, separator, and decoder. The mixture encoder transforms segments of the binaural mixture signal to an intermediate feature representation. The separator – guided by the speaker embedding obtained from reference speech of the target speaker – estimates weights that are applied to the intermediate feature representation using a weight sum method. The decoder reconstructs the (binaural) target speaker signals from the weighted intermediate feature representation. Both systems fulfil the causality constraint of the challenge, i.e., no information from input samples more than 5 ms in the future is used to obtain an output sample (see Table 1).

2.1. Temporal Convolutional Network (TCN)-based System

Fig. 2 depicts the proposed TCN-based system, where the speaker separator network is based on 2 stacks of TCN blocks with 8 layers each. Each TCN layer [4] consists of two 1-dimensional convolutional (1D-CNN) layers, two parametric rectified linear unit (PReLU) activation functions, and one dilated depth-wise separable convolutional layer. The speaker embedder network of this system utilizes a single TCN block. To incorporate the speaker embedding, the feature-wise linear modulation (FiLM) fusion method [5] is utilized, which applies an affine transformation to the TCN residual activations. The mixture encoder consists of a 1D-CNN layer taking all 6 channels of the hearing aids as input, while the decoder consists of a transposed 1D-CNN layer producing binaural output signals.

2.2. TCN-Conformer-based System

Fig. 1 depicts the proposed TCN-Conformer-based system which is a binaural extension of [1]. The speaker separator network is based on 4 stacks of TCN blocks and conformer blocks [6], where each TCN block is followed by a conformer block. Each TCN block [4] consists of two 1D-CNN layers, two PReLU activations and one dilated depth-wise separable convolutional layer. Each conformer block consists of four different blocks: two feed-forward blocks, one multi-head self-attention block, and a convolutional block. The multi-head attention block is utilized between the first feed-forward block and convolutional block. The second feed-forward block is applied just after the convolutional block. The speaker embedder network of this system utilizes a ResNet-based architecture [3]. To incorporate the speaker embedding, the concatenation method is utilized, where the speaker embedding is repeatedly concatenated with the outputs of the conformer blocks along the feature dimension. The input to the first TCN block is the concatenation of the encoded features obtained from the mixture signal using the mixture encoder and the speaker embedding estimated using the speaker embedder network, while the inputs to the other TCN blocks are the concatenation of the conformer block output and the speaker embedding.

3. Experiments

3.1. Dataset

Before processing, the microphone signals were downsampled from 44.1 kHz to 16 kHz. The official training and development datasets (extended by a factor 10 using the official CEC2-provided tools) were used to train and validate both proposed systems, i.e., no external data or augmentation techniques were used. Furthermore, none of our submitted systems uses the provided head rotation signals.

3.2. Training settings

For all experiments, speaker embeddings are generated with randomly chosen utterances of the target speaker different from the mixture utterance. Table 1 shows the parameters of the mixture encoder and decoder, speaker embedder, and TCN layers for both proposed systems. Additionally, the conformer blocks of TCN-Conformer-based system utilize 8-head attention, while the convolutional kernel size is fixed to 31. The output of the first convolution layer is expanded with a factor 3 in each block, while the output of the linear layer is set to be 4 times the input size. A weighted combination of the scale-invariant signal-to-noise ratio (SI-SNR) loss [7] and the cross-entropy loss was used to train the TCN-Conformer-based system, whereas a weighted combination of the mean spectral absolute error loss [8] and the cross-entropy loss was used to train the TCN-based system. For both systems, the reverberant target speech at CH1 was used as the binaural target signal. Both systems were trained using early stopping criteria.

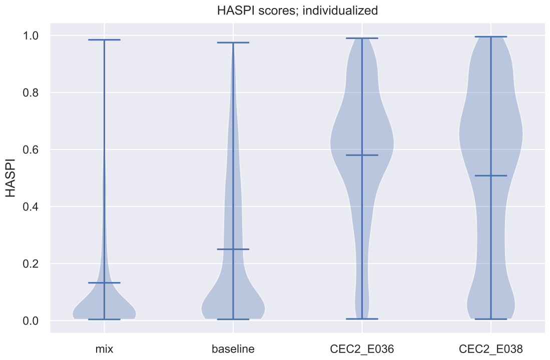

Figure 3: *HASPI scores obtained for mixture, baseline, and proposed systems on the last 500 utterances of the development dataset.*

Table 1: *Parameter settings used for the submitted systems.*

	CEC2_E036	**CEC2_E038**
Frame length	2.5 ms	2 ms
Frame shift	1.25 ms	1 ms
Encoder/Decoder (dim)	128	512
Speaker embedding (dim)	256	512
Kernel-size	3	3
Conv-channel	256	512
Conv-type	causal	causal
Masking technique	weight & sum [9]	weight & sum [9]

4. Results

In this section, we present the HASPI evaluation results based on a subset of the CEC2 development dataset consisting of the last 500 utterances. For each utterance, audiograms detailing the hearing loss of three specific listeners were provided, which were used in the hearing loss compensation stage to individualize the output signals for the specific listeners. Figure 3 depicts a violin plot of the HASPI scores obtained for the mixture, the baseline system, and both submitted systems. It can be observed that both submitted systems achieve a significant improvement compared to the mixture as well as the baseline system. The mean HASPI scores are equal to 0.13 (mixture), 0.25 (baseline), 0.58 (CEC_036), and 0.51 (CEC_038). It is interesting to note that the distribution of the HASPI scores is quite different for both submitted systems, which can presumably be explained by a different performance at (very) low SNRs.

5. Conclusion

Aiming to improve speech intelligibility for hearing-impaired listeners in a reverberant scenario with a target speaker and multiple interfering sources, in our contribution we proposed two DNN-based speaker-informed target speaker extraction systems for the CEC2. Experimental results obtained on a subset of the official development dataset demonstrate the advantage of both proposed systems compared with the baseline system.

6. References

[1] R. Sinha, M. Tammen, C. Rollwage, and S. Doclo, "Speaker-conditioning Single-channel Target Speaker Extraction using Conformer-based Architectures," May 2022.

[2] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," in *Proc. Interspeech*, Graz, Austria, 2019, pp. 2728–2732.

[3] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Spex+: A complete time domain speaker extraction network," in *Proc. Interspeech*, Shanghai, China, 2020, pp. 1406–1410.

[4] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[5] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual Reasoning with a General Conditioning Layer," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018.

[6] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[7] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?" in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Brighton, UK, May 2019, pp. 626–630.

[8] Z.-Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust asr," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 1778–1787, 2020.

[9] C. Han, Y. Luo, and N. Mesgarani, "Binaural speech separation of moving speakers with preserved spatial cues." in *Interspeech*, 2021, pp. 3505–3509.