Combining binaural LCMP beamforming and deep multi-frame filtering for joint dereverberation and interferer reduction in the Clarity-2021 Challenge

Marvin Tammen, Henri Gode, Hendrik Kayser, Eike J. Nustede, Nils L. Westhausen, Jörn Anemüller, Simon Doclo

Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, University of Oldenburg, Germany

marvin.tammen@uol.de

Abstract

In this paper we present our algorithms submitted to the Clarity-2021 Challenge [1], aiming at improving speech intelligibility for hearing-impaired listeners in a reverberant acoustic scenario with a target speaker and an interfering speaker. The algorithms consist of a weighted binaural linearly-constrained-minimum-power beamformer, performing simultaneous dereverberation and interferer reduction, a deep binaural multi-frame filter to reduce residual interference, and a dynamic range compression stage for audiogram-based hearing loss compensation. For all submitted systems the MBSTOI results indicate a significant improvement compared with the baseline system.

1. Algorithm description

Figure 1 depicts the block diagram of the proposed algorithms, consisting of a binaural beamformer (see Section 1.1), an optional deep learning-based post-processing stage (see Section 1.2) and dynamic range compression (see Section 1.3). The combination of these algorithmic blocks into the three systems submitted to the challenge will be explained in more detail in Section 2. Before processing, the microphone signals have been resampled from 44.1 kHz to 16 kHz.

1.1. Weighted binaural LCMP beamformer

Aiming at preserving the target speaker, reducing the interfering speaker and preserving the binaural cues of both speakers, we used an adaptive version of the weighted binaural linearly-constrained-minimum-power (wBLCMP) beamformer proposed in [2]. The wBLCMP beamformer unifies weighted prediction error (WPE) dereverberation and binaural LCMP beamforming [3, 4] to simultaneously perform dereverberation and interferer reduction. Similarly as in [5], the convolutional beamformer is optimized using a sparsity-promoting ℓ_p-norm cost function, leading to an iterative reweighted least squares (IRLS) algorithm. In each iteration, the $M(L_h-\tau+1)\times 2$-dimensional convolutional binaural beamformer $\mathbf{H}_t$, with t the time frame index, M the number of microphones ($M=6$), L_h the filter length and τ the prediction delay, is given in each STFT frequency bin as

$$\mathbf{H}_t = \mathbf{R}_t^{-1}\mathbf{C}_t\left[\mathbf{C}_t^{\mathrm{H}}\mathbf{R}_t^{-1}\mathbf{C}_t\right]^{-1}\begin{bmatrix}1 & 0\\ 0 & \delta\end{bmatrix}\mathbf{C}_t^{\mathrm{H}}[\mathbf{e}_L,\mathbf{e}_R], \quad (1)$$

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project ID 352015383 (SFB 1330 B2 and B3) and Project ID 390895286 (EXC 2177/1). Research reported in this publication was supported by the National Institute On Deafness And Other Communication Disorders of the National Institutes of Health under Award Number R01DC015429. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

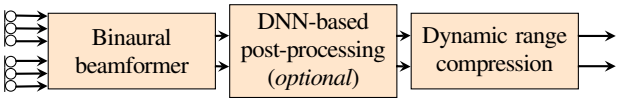

Figure 1: *Block diagram of the proposed algorithms, consisting of a weighted binaural LCMP beamformer, an optional deep learning-based post-processing stage (deep binaural MFMVDR filter) and dynamic range compression.*

where $\mathbf{R}_t$ is a weighted covariance matrix of the microphone signals $\bar{\mathbf{y}}_t = \left[\mathbf{y}_t^{\mathrm{T}}\ \mathbf{y}_{t-\tau}^{\mathrm{T}}\ \cdots\ \mathbf{y}_{t-L_h+1}^{\mathrm{T}}\right]^{\mathrm{T}}$, $\mathbf{C}_t$ contains the relative transfer functions (RTFs) of the target speaker and the interfering speaker, δ is a parameter determining the amount of interferer reduction, and $\mathbf{e}_L$ and $\mathbf{e}_R$ are selection vectors corresponding to the left and right frontal microphones on the hearing aids. The RTF of the interfering speaker is computed as the normalized principal eigenvector of the covariance matrix estimated during the first 2 seconds (only interferer active), whereas the RTF of the target speaker is adaptively estimated using the covariance whitening method [6] after 2 seconds (target and interferer active).

For the STFT framework we used a frame length of 80 samples (corresponding to 5ms), a square-root Hann window, and a frame shift of 40 samples in a weighted-overlap-add processing scheme. We used the following parameters: filter length $L_h=8$, prediction delay $\tau=2$, shape parameter $p=0.5$, and interferer reduction parameter $\delta=0.1$.

1.2. Deep binaural MFMVDR filter

Aiming at reducing residual interference at the output of the wBLCMP beamformer while preserving the correlated speech components, we used a binaural extension of the deep multi-frame minimum-variance-distortionless-response (MFMVDR) filter proposed in [7], termed deep binaural MFMVDR (BMFMVDR) filter. Similarly to [7], the required parameters of the BMFMVDR filter, i.e., the covariance matrices and the speech interframe correlation vectors, are estimated by minimizing the scale-dependent signal-to-distortion-ratio [8] loss function at the output of the BMFMVDR filter using causal temporal convolutional networks (TCNs). A PyTorch implementation of the BMFMVDR filter will be made publicly available.

For the STFT framework we used the same parameters as for the wBLCMP beamformer. The deep BMFMVDR filter used a filter length of 4, and it was trained on the official Clarity-2021 Challenge training data for 67 epochs using the AdamW optimizer with an initial learning rate of 10^{-3} (which was halved after 3 consecutive epochs without validation loss improvement), a weight decay of 10^{-2}, and a batch size of 4 using an NVIDIA GeForce®

RTX 3090 graphics card. For the employed TCNs, we used 2 stacks of 8 layers each, with a kernel size of 3, resulting in a temporal receptive field of about 2.56s and 3.02M parameters.

1.3. Dynamic range compression

The dynamic range compression (DRC) stage is used for audiogram-based compensation of hearing loss and further level adjustments. It consists of a spectral-domain multi-band dynamic range compressor (MBDRC) that implements a noise gate, frequency- and hearing-loss-dependent amplification and limitation of the maximum output level, and a volume control at the output. As an alternative to MBDRC, the "half-gain rule" (HGR) was used for hearing loss compensation, i.e., only volume control was applied set to the pure-tone average of 500 Hz, 1000 Hz, and 2000 Hz divided by 2. The system also takes care of calibration and soft-clipping of the output audio signal, with settings adopted from the challenge baseline system. The STFT and filterbank parameters and the noise gate levels for the MBDRC were adopted from the challenge baseline system. The gains applied in the MBDRC were computed using the compressive *Camfit* gain prescription rule [9].

2. Submitted Systems

All submitted systems use the wBLCMP beamformer (Section 1.1) as first processing stage and dynamic range compression (Section 1.3) as last processing stage. The third submission system uses an additional deep learning-based post-processing stage after the wBLCMP beamformer and before the dynamic range compression stage.

- **CEC1_E016**: Combination of wBLCMP beamformer and HGR-based hearing loss compensation.
- **CEC1_E019**: Combination of wBLCMP beamformer and MBDRC.
- **CEC1_E021**: Combination of wBLCMP beamformer, deep BMFMVDR filter and MBDRC.

For the DRC stage, the parameters in Table 1 were selected for each of the submitted systems based on the results obtained on a small development data subset: output gain $\mathrm{vol}_{\mathrm{out}}$, MBDRC maximum output level $\mathrm{lev}_{\mathrm{max}}$, attack time τ_{att} and decay time τ_{dec} of the MBDRC, as well as soft-clipping threshold $\mathrm{sc}_{\mathrm{thr}}$.

Table 1: *Parameter values used in the DRC stage for the submitted systems.*

	CEC1_E016	**CEC1_E019**	**CEC1_E021**
$\mathrm{vol}_{\mathrm{out}}$ (dB)	HGR	10	10
$\mathrm{lev}_{\mathrm{max}}$ (dB)	—	120	120
τ_{att} (s)	—	0.002	0.001
τ_{dec} (s)	—	0.01	0.01
$\mathrm{sc}_{\mathrm{thr}}$ (dB)	117	117	117

3. Results

In this section, we present the simulation results based on the development dataset. For the evaluation, we first processed the output signals of our submitted systems using the provided code of the hearing loss model, before estimating the speech intelligibility using the provided MBSTOI measure. Figure 2 depicts the MBSTOI results for the baseline system and the three submitted systems. It can be observed that all submitted systems achieve a significant improvement compared with the baseline system. Furthermore, the differences between the submitted systems in terms of MBSTOI are relatively small, indicating that neither the more sophisticated MBDRC hearing loss compensation nor the DNN-based post-processing stage achieve a significant improvement in terms of speech intelligibility upon the system CEC1_E016. Nevertheless, since the output signals of the submitted systems sounded quite differently with respect to interferer reduction, artefacts and high-frequency content, we decided to submit all three systems.

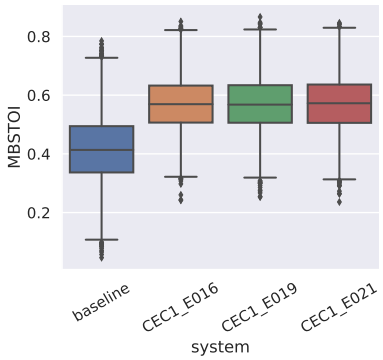

Figure 2: *MBSTOI results of the baseline system and the submitted systems on the development dataset.*

4. References

[1] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, and R. Viveros Muñoz, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in *Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Brno, Czech Republic, 2021.

[2] A. Aroudi, M. Delcroix, T. Nakatani, K. Kinoshita, S. Araki, and S. Doclo, "Cognitive-Driven Convolutional Beamforming Using EEG-Based Auditory Attention Decoding," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing*, Espoo, Finland, Sep. 2020, pp. 1–6.

[3] E. Hadad, S. Doclo, and S. Gannot, "The Binaural LCMV Beamformer and its Performance Analysis," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 543–558, Mar. 2016.

[4] N. Gößling, D. Marquardt, I. Merks, T. Zhang, and S. Doclo, "Optimal binaural LCMV beamforming in complex acoustic scenarios: Theoretical and practical insights," in *Proc. International Workshop on Acoustic Signal Enhancement*, Tokyo, Japan, 2018, pp. 381–385.

[5] H. Gode, M. Tammen, and S. Doclo, "Joint multi-channel dereverberation and noise reduction using a unified convolutional beamformer with sparse priors," *arXiv:2106.01902*, 2021.

[6] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, 2009.

[7] M. Tammen and S. Doclo, "Deep multi-frame MVDR filtering for single-microphone speech enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, Canada, June 2021, pp. 8443–8447.

[8] 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.

[9] B. Moore, J. Alcántara, M. Stone, and B. Glasberg, "Use of a loudness model for hearing aid fitting: II. Hearing aids with multi-channel compression," *British journal of audiology*, vol. 33, no. 3, pp. 157–170, 1999.