Combining weighted 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 Enhancement Challenge, aiming at improving speech intelligibility for hearing-impaired listeners in a reverberant acoustic scenario with a target speaker and an interfering source. The algorithms combine 1) a weighted binaural linearly constrained minimum power beamformer, performing joint dereverberation and interferer reduction, 2) a deep binaural multi-frame postfilter to reduce residual interference, and 3) an audiogram-based hearing loss compensation stage. Objective metrics as well as subjective listening experiments with hearing-impaired listeners show that all submitted systems result in a significant improvement in terms of speech intelligibility compared with the baseline system.

1. Introduction

In the Clarity Enhancement Challenge (CEC1) [1], a hearing-impaired listener is considered at a fixed position and orientation in a moderately reverberant room, wearing 2 hearing aids with 3 microphones on each hearing aid. Two competing sources, i.e. one target speaker and one interfering source (speech or noise), are simulated with stationary room impulse responses (RIRs), where the target speaker starts 2s after the interfering source. In general, acoustic disturbances, such as interfering sources, ambient noise, reverberation, as well as hearing loss are known to degrade speech intelligibility [2–5]. We propose to tackle this challenging scenario with a system consisting of 3 cascaded blocks (see Figure 1), i.e., a binaural beamformer for joint interferer reduction and dereverberation, a binaural postfilter for residual interferer reduction, and a hearing loss compensation stage. In the following subsections, we will first provide some general background information about these three blocks, before presenting the details of the proposed system in Section 2.

1.1. Binaural Beamforming

Different multi-microphone techniques have been proposed in the literature to reduce noise, interfering sources and/or reverberation [6, 7]. A commonly used multi-microphone noise reduction technique is the minimum power distortionless response (MPDR) beamformer [6, 8], which aims at minimizing the output power while leaving the desired speech component undistorted. The linearly constrained minimum power (LCMP) beamformer generalizes the MPDR beamformer, providing the possibility of multiple linear constraints, e.g., to perform controlled interferer reduction [8–10]. Often the constraints are formulated in terms of the relative transfer function (RTF) vectors of the sources [11].

To achieve dereverberation, the weighted prediction error (WPE) technique and its variants are commonly employed [12–14]. WPE uses a convolutional filter, applied to a number of past frames in the short-time Fourier transform (STFT) domain, to estimate and subtract the late reverberation component. Since the usual WPE cost functions do not have analytic solutions, it has been proposed to use iterative alternating optimization schemes. Aiming at joint dereverberation and noise reduction, it has been proposed to perform multiple-input multiple-output (MIMO) WPE as a preprocessing stage before MPDR beamforming in a cascade system [15]. By unifying the optimization of the convolutional WPE filter and the MPDR beamformer the so-called weighted power minimization distortionless response (WPD) convolutional beamformer [16, 17] and its generalization using sparse priors [18] were shown to outperform cascade systems. The unified WPD beamformer is optimized similarly to the WPE filter with an additional distortionless constraint using the RTF of the target speaker.

Aiming at jointly performing dereverberation and interferer reduction and preserving the binaural cues of all sources, the weighted binaural linearly constrained minimum power (wBLCMP) beamformer proposed in [19] generalizes the WPD beamformer by unifying WPE dereverberation and LCMP beamforming [9, 10]. In our CEC1 contribution we used an adaptive version of this wBLCMP beamformer. Similarly as in [18], the convolutional beamformer is computed by minimizing a sparsity-promoting ℓ_p-norm cost function.

1.2. Binaural Postfilter

Typically, some residual interference will remain in the output of the wBLCMP beamformer. While this may be desirable, e.g., to preserve awareness of the acoustic scene, it may also be desirable to achieve more interferer reduction. A common approach to achieve this is to include a postfilter at the output of the beamformer [6, 7]. Noting that the goal is to perform interferer reduction in a binaural listening scenario, an additional desired property of the postfilter is to preserve the binaural cues of the target speaker.

With these considerations in mind, in our CEC1 contribution we propose a binaural extension of the deep multi-frame minimum variance distortionless response (MFMVDR) filter [20], termed deep binaural MFMVDR (BMFMVDR) filter, as the postfilter. The deep BMFMVDR filter aims at minimizing the power spectral density of the undesired components while preserving the binaural correlated components of the target speaker (and, as a result, the corresponding binaural cues). Similarly as in [20], all required parameters, i.e., the noisy and undesired spatio-temporal covariance

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.

matrices as well as the inter-frame correlation vectors, are estimated by minimizing the scale-dependent signal-to-distortion ratio [21] using temporal convolutional networks [22].

1.3. Hearing Loss Compensation

To compensate for the potential hearing loss of the listener, we consider two options, which both make use of the bilateral pure-tone audiograms. The first option consists of a multi-band dynamic range compressor [23], which is also used in the CEC1 baseline system. The second option consists of a simple broadband gain which is computed based on the half-gain rule.

2. Proposed System

Figure 1 depicts a block diagram of the proposed algorithms, consisting of a weighted binaural LCMP beamformer (see Section 2.1), an optional deep binaural MFMVDR postfilter (see Section 2.2) and a hearing loss compensation stage (see Section 2.3). The combination of these algorithmic blocks into the three systems submitted to the challenge will be explained in more detail in Section 3.

2.1. Weighted binaural LCMP beamformer

2.1.1. Signal Model

Although only 2 sources and no ambient noise are present in the CEC1 scenario, we will explain the wBLCMP beamformer for a more general acoustic scenario, consisting of J sources captured by M microphones in a noisy environment. Without loss of generality, the first source is considered to be the target speaker. The STFT coefficients of the microphone signals at time frame t and any frequency bin are given by

$$\mathbf{y}_t = \begin{bmatrix} y_{1,t} & \cdots & y_{M,t} \end{bmatrix}^{\mathrm{T}} \in \mathbb{C}^{M\times 1}, \tag{1}$$

with $(\cdot)^{\mathrm{T}}$ denoting the transpose operator. The frequency index is omitted for brevity since it is assumed that each frequency subband is independent and hence can be processed individually. Similarly to [16–18], the multi-channel microphone signal $\mathbf{y}_t$ is modeled as the sum of the convolutions of each source signal $s_{j,t}$ with its respective multi-channel convolutive transfer function (CTF) matrix $\mathbf{A}_j = \begin{bmatrix} \mathbf{a}_{j,0} & \cdots & \mathbf{a}_{j,L_a-1} \end{bmatrix} \in \mathbb{C}^{M\times L_a}$ plus additive noise $\mathbf{n}_t \in \mathbb{C}^{M\times 1}$, i.e.

$$\mathbf{y}_t = \sum_{j=1}^{J}\sum_{l=0}^{L_a-1} \mathbf{a}_{j,l} s_{j,t-l} + \mathbf{n}_t \tag{2}$$

$$= \sum_{j=1}^{J}\underbrace{\sum_{l=0}^{\tau-1} \mathbf{a}_{j,l} s_{j,t-l}}_{:=\mathbf{d}_{j,t}} + \sum_{j=1}^{J}\underbrace{\sum_{l=\tau}^{L_a-1} \mathbf{a}_{j,l} s_{j,t-l}}_{:=\mathbf{r}_{j,t}} + \mathbf{n}_t, \tag{3}$$

where L_a denotes the number of taps of the CTFs. The delay τ separates the early reflections from the late reverberation, decomposing the reverberant signal for the j-th source into its direct component $\mathbf{d}_{j,t} \in \mathbb{C}^{M\times 1}$ (including early reflections) and its late reverberation component $\mathbf{r}_{j,t} \in \mathbb{C}^{M\times 1}$. The direct component for the j-th source can be approximated using the multiplicative transfer function (MTF) vector $\mathbf{v}_{j,t} \in \mathbb{C}^{M\times 1}$ as [24]

$$\mathbf{d}_{j,t} \approx \mathbf{v}_{j,t} s_{j,t} = \tilde{\mathbf{v}}_{j,m,t} d_{j,m,t} \quad j\in\{1,...,J\}, \quad m\in\{1,...,M\}, \tag{4}$$

where $d_{j,m,t}$ denotes the direct component of the j-th source in the reference microphone m at time frame t. The vector $\tilde{\mathbf{v}}_{j,m,t} = \mathbf{v}_{j,t}/v_{j,m,t} \in \mathbb{C}^{M\times 1}$ denotes the RTF vector for the j-th source, where $v_{j,m,t}$ is the m-th entry of $\mathbf{v}_{j,t}$.

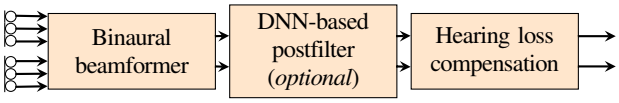

Figure 1: *Block diagram of the proposed algorithms, consisting of a weighted binaural LCMP beamformer, an optional deep binaural MFMVDR postfilter, and a hearing loss compensation stage.*

2.1.2. Convolutional Filter

To obtain an estimate $z_{m,t}$ of the direct target speech component $d_{1,m,t}$ in the reference microphone m at time frame t, it has been proposed in [16–19] to apply a convolutional filter $\bar{\mathbf{h}}_{m,t} \in \mathbb{C}^{M(L_h-\tau+1)\times 1}$ to the noisy STFT vector in (1), i.e.

$$z_{m,t} = \bar{\mathbf{h}}_{m,t}^{\mathrm{H}} \bar{\mathbf{y}}_t, \tag{5}$$

where $(\cdot)^{\mathrm{H}}$ denotes the conjugate transpose operator and the stacked microphone signal vector $\bar{\mathbf{y}}_t$ is defined as

$$\bar{\mathbf{y}}_t = \begin{bmatrix} \mathbf{y}_t^{\mathrm{T}} & | & \mathbf{y}_{t-\tau}^{\mathrm{T}} & \cdots & \mathbf{y}_{t-L_h+1}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \in \mathbb{C}^{M(L_h-\tau+1)\times 1}. \tag{6}$$

It should be noted that the vector $\bar{\mathbf{y}}_t$ only includes a subset of the L_h most recent frames, i.e. it includes the current frame but excludes the preceding $\tau-1$ frames, aiming at preserving the early reflections.

2.1.3. Filter Optimization

We propose to optimize the convolutional wBLCMP beamformer by explicitly taking into account that the direct target speech component in the STFT domain is sparser than the noisy reverberant mixture recorded by the microphones. Hence similarly to the WPD variant in [18], we propose to optimize the convolutional filter in (5) using an ℓ_p-norm cost function, i.e.

$$\mathcal{L}(\bar{\mathbf{h}}_{m,t}) \propto \sum_{n=1}^{t} \gamma^{t-n} |z_{m,n}|^p, \tag{7}$$

where $p \in (0,2]$ denotes the so-called shape parameter and the smoothing parameter $\gamma \in (0,1]$ allows adaptation to possibly time-varying transfer functions. The shape parameter determines the sparsity of the cost function, where small values of p promote sparsity and it should be noted that for $0<p<1$ this cost function is non-convex. In addition, linear constraints for each source are imposed using their RTFs, i.e.

$$\bar{\mathbf{h}}_{m,t}^{\mathrm{H}} \bar{\mathbf{v}}_{j,m,t} = \beta_j, \tag{8}$$

where β_j denotes a scaling factor for each source and $\bar{\mathbf{v}}_{j,m,t} = \begin{bmatrix} \tilde{\mathbf{v}}_{j,m,t}^{\mathrm{T}} & \mathbf{0}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$ with $\mathbf{0}$ a vector containing $M(L_h-\tau)$ zeros. The scaling factor β_1 is usually set to 1, corresponding to a distortionless constraint for the target speaker, whereas all other scaling factors are usually chosen to be close to 0, aiming at suppressing the interfering sources.

Similarly as in [13, 18, 25], we propose to use an iteratively reweighted least squares (IRLS) method to minimize the cost function in (7) subject to the constraints in (8). The basic idea is to replace the non-convex ℓ_p-norm minimization problem with a series of convex ℓ_2-norm minimization subproblems, where in each iteration the ℓ_2-norm minimization subproblem has an analytic solution. The two alternating steps are described in the following paragraphs.

(1) Constrained ℓ_2–Norm Subproblem Minimization In each iteration i, the non-convex cost function in (7) is replaced with a convex weighted ℓ_2-norm cost function, i.e.

$$\mathcal{L}\big(\bar{\mathbf{h}}_{m,t,i}\big) \propto \sum_{n=1}^{t} \gamma^{t-n} w_{n,i} |z_{m,n,i}|^p \tag{9}$$

where the weights $w_{n,i}$ are real-valued and positive. The binaural filter minimizing (9) subject to the linear constraints in (8) with respect to the left and right reference microphone denoted by $m = \nu \in \{L,R\}$ is equal to

$$\boxed{\bar{\mathbf{h}}_{\nu,t,i} = \bar{\mathbf{R}}_{y,t,i}^{-1} \bar{\mathbf{C}}_t \left(\bar{\mathbf{C}}_t^{\mathrm{H}} \bar{\mathbf{R}}_{y,t,i}^{-1} \bar{\mathbf{C}}_t \right)^{-1} \mathbf{B} \bar{\mathbf{C}}_t^{\mathrm{H}} \mathbf{e}_\nu} \tag{10}$$

where $\bar{\mathbf{R}}_{y,t,i} = \sum_{n=1}^{t} \gamma^{t-n} w_{n,i} \bar{\mathbf{y}}_n \bar{\mathbf{y}}_n^{\mathrm{H}}$ denotes the weighted noisy spatio-temporal covariance matrix of the stacked microphone signals, $\bar{\mathbf{C}}_t = \begin{bmatrix} \bar{\mathbf{v}}_{1,\nu,t} & \cdots & \bar{\mathbf{v}}_{J,\nu,t} \end{bmatrix}$ denotes the constraint matrix containing the RTF vectors for all sources, $\mathbf{B} = \mathrm{diag}\left(\begin{bmatrix} \beta_1 & \cdots & \beta_J \end{bmatrix}^{\mathrm{T}} \right)$ denotes the diagonal scaling matrix containing the scaling factors for all sources, and $\mathbf{e}_\nu$ is a selection vector with its entry corresponding to the left or right reference microphone equal to 1 and all other entries equal to 0.

(2) Weight Update Similarly as in [13, 18, 25], in each iteration the weights in (9) are updated as

$$w_{t,i+1} = \left(\sum_{\nu} |z_{\nu,t,i}|^2 \right)^{p/2-1} = \left(\sum_{\nu} |\bar{\mathbf{h}}_{\nu,t,i} \bar{\mathbf{y}}_t|^2 \right)^{p/2-1}, \tag{11}$$

such that (9) is a first-order approximation of (7).

2.1.4. Parameter Estimation

The wBLCMP beamformer in (10) requires an estimate of the RTFs of each source. In the CEC1 scenario, one stationary target speaker ($j = 1$) and one stationary interfering source ($j = 2$) are present, where the target speaker starts exactly 2s after the interfering source. Taking advantage of this scenario allows to estimate the (stationary) RTF of the interfering source as the normalized principal eigenvector of the interference covariance matrix $\mathbf{R}_i = \sum_{n=1}^{t \triangleq 2\mathrm{s}} \mathbf{y}_n \mathbf{y}_n^{\mathrm{H}}$. The RTF of the target speaker can then be estimated using the covariance whitening method [26], i.e. based on the generalized eigenvalue decomposition of the noisy covariance matrix $\mathbf{R}_{y,t} = \sum_{n \triangleq 2\mathrm{s}}^{t} \mathbf{y}_n \mathbf{y}_n^{\mathrm{H}}$ and the interference covariance matrix $\mathbf{R}_i$. The RTF of the target speaker is constantly updated for every frame t starting from 2s.

2.2. Deep Binaural Multi-Frame MVDR Filter

In this section, we describe the proposed deep BMFMVDR filter, which is used as the binaural postfilter of the wBLCMP beamformer in one of our submitted systems. The deep BMFMVDR filter is a binaural extension of the deep MFMVDR filter proposed in [20].

2.2.1. Signal Model

We consider the binaural output of the wBLCMP beamformer in the STFT domain $\mathbf{z}_t = [z_{L,t},\, z_{R,t}]^T = \mathbf{x}_t + \mathbf{i}_t$ as the binaural input signal of the deep BMFMVDR filter, where $\mathbf{x}_t$ and $\mathbf{i}_t$ denote the residual speech and interference components.

Stacking the N most recent time frames in a vector, we obtain the binaural multi-frame signal vector $\bar{\mathbf{z}}_t = \begin{bmatrix} z_{L,t} & \cdots & z_{L,t-N+1} & z_{R,t} & \cdots & z_{R,t-N+1} \end{bmatrix}^T = \bar{\mathbf{x}}_t + \bar{\mathbf{i}}_t$, with the vectors $\bar{\mathbf{x}}_t$ and $\bar{\mathbf{i}}_t$ defined similarly.

In [27], it has been proposed to exploit the speech correlation across adjacent STFT frames by decomposing the (single-microphone) multi-frame speech vector into a temporally correlated and a temporally uncorrelated part. Similarly, in a binaural scenario, the binaural multi-frame speech vector $\bar{\mathbf{x}}_t$ can be decomposed into a spatio-temporally correlated and a spatio-temporally uncorrelated part w.r.t. the left or the right speech component $x_{L,t}$ or $x_{R,t}$, i.e.,

$$\bar{\mathbf{x}}_t = \underbrace{\boldsymbol{\gamma}_{\mathbf{x},\nu,t} x_{\nu,t}}_{\text{correlated}} + \underbrace{\bar{\mathbf{x}}'_{\nu,t}}_{\text{uncorrelated}}, \tag{12}$$

with $\nu \in \{L,R\}$ denoting the left or right channel. The (highly time-varying) left or right speech inter-frame correlation (IFC) vector $\boldsymbol{\gamma}_{x,\nu,t}$ describes the correlation between the current and previous time frames w.r.t. the left or right speech STFT coefficient $x_{\nu,t}$.

The binaural speech component $\mathbf{x}_t$ is estimated by applying (complex-valued) finite impulse response filters $\mathbf{w}_{L,t}$ and $\mathbf{w}_{R,t}$ with $2N$ taps each to the binaural multi-frame signal vector, i.e.,

$$\widehat{\mathbf{x}}_t = \begin{bmatrix} \mathbf{w}_{L,t}^{\mathrm{H}} \bar{\mathbf{z}}_t \\ \mathbf{w}_{R,t}^{\mathrm{H}} \bar{\mathbf{z}}_t \end{bmatrix}. \tag{13}$$

Assuming that the speech and interference components are uncorrelated, the $2N \times 2N$-dimensional input spatio-temporal covariance matrix (STCM) $\boldsymbol{\Phi}_{z,t} = \mathcal{E}\left\{ \bar{\mathbf{z}}_t \bar{\mathbf{z}}_t^{\mathrm{H}} \right\}$, with $\mathcal{E}\{\cdot\}$ the expectation operator, can be written as

$$\boldsymbol{\Phi}_{z,t} = \boldsymbol{\Phi}_{x,t} + \boldsymbol{\Phi}_{i,t}, \tag{14}$$

with the speech and interference STCMs $\boldsymbol{\Phi}_{x,t} = \mathcal{E}\left\{ \bar{\mathbf{x}}_t \bar{\mathbf{x}}_t^{\mathrm{H}} \right\}$ and $\boldsymbol{\Phi}_{i,t} = \mathcal{E}\left\{ \bar{\mathbf{i}}_t \bar{\mathbf{i}}_t^{\mathrm{H}} \right\}$. Using (12), the STCM in (14) can be rewritten as

$$\boldsymbol{\Phi}_{z,t} = \phi_{x,\nu,t} \boldsymbol{\gamma}_{x,\nu,t} \boldsymbol{\gamma}_{x,\nu,t}^{\mathrm{H}} + \underbrace{\boldsymbol{\Phi}_{x',\nu,t} + \boldsymbol{\Phi}_{i,t}}_{=: \boldsymbol{\Phi}_{u,\nu,t}}, \tag{15}$$

where the undesired STCM $\boldsymbol{\Phi}_{u,\nu,t}$ consists of both the uncorrelated speech STCM $\boldsymbol{\Phi}_{x',\nu,t}$ as well as the interference STCM $\boldsymbol{\Phi}_{i,t}$. Using (15), the speech IFC vector $\boldsymbol{\gamma}_{x,\nu,t}$ can be written as

$$\boldsymbol{\gamma}_{x,\nu,t} = \frac{1 + \xi_{\nu,t}}{\xi_{\nu,t}} \frac{\boldsymbol{\Phi}_{z,t} \mathbf{e}_\nu}{\mathbf{e}_\nu^{\mathrm{T}} \boldsymbol{\Phi}_{z,t} \mathbf{e}_\nu} - \frac{1}{\xi_{\nu,t}} \frac{\boldsymbol{\Phi}_{u,\nu,t} \mathbf{e}_\nu}{\mathbf{e}^{\mathrm{T}} \boldsymbol{\Phi}_{u,\nu,t} \mathbf{e}_\nu}, \tag{16}$$

where $\xi_{\nu,t} = \frac{\mathbf{e}_\nu^{\mathrm{T}} \boldsymbol{\Phi}_{x,\nu,t} \mathbf{e}_\nu}{\mathbf{e}_\nu^{\mathrm{T}} \boldsymbol{\Phi}_{u,\nu,t} \mathbf{e}_\nu}$ denotes the a-priori signal-to-noise ratio (SNR), and $\mathbf{e}_\nu$ is a selection vector with its first or $(N+1)$-th entry equal to 1 and all other entries equal to 0.

2.2.2. Optimization Problem and Solution

In [27], the MFMVDR filter for single-microphone speech enhancement has been proposed, aiming at minimizing the output undesired power spectral density (PSD) while not distorting the correlated speech component. In our CEC1 contribution, we propose to extend the MFMVDR filter to a binaural filter by considering the left and right channels independently, i.e.,

$$\underset{\mathbf{w}_{\nu,t}}{\mathrm{argmin}} \quad \mathbf{w}_{\nu,t}^H \boldsymbol{\Phi}_{u,\nu,t} \mathbf{w}_{\nu,t}, \quad \text{s.t.} \quad \mathbf{w}_{\nu,t}^H \boldsymbol{\gamma}_{x,\nu,t} = 1. \tag{17}$$

Solving this constrained optimization problem yields the BMFMVDR filter vectors:

$$\boxed{\mathbf{w}_{\nu,t}^{\mathrm{BMFMVDR}} = \frac{\boldsymbol{\Phi}_{u,\nu,t}^{-1} \boldsymbol{\gamma}_{x,\nu,t}}{\boldsymbol{\gamma}_{x,\nu,t}^H \boldsymbol{\Phi}_{u,\nu,t}^{-1} \boldsymbol{\gamma}_{x,\nu,t}}} \tag{18}$$

2.2.3. Parameter Estimation

To compute the BMFMVDR filters in (18), we require estimates of the undesired STCMs $\mathbf{\Phi}_{u,\nu,t}$ as well as the speech IFC vectors $\boldsymbol{\gamma}_{x,\nu,t}$. Rather than estimating $\boldsymbol{\gamma}_{x,\nu,t}$ directly, we use the indirect estimator in (16), resulting in the need to estimate the input STCM $\mathbf{\Phi}_{z,t}$ as well as the binaural a-priori SNRs $\xi_{\nu,t}$.

Similarly as in [20], we propose to estimate the required parameters from the binaural input signals $\mathbf{z}_t$ using a deep learning-based approach by minimizing the scale-dependent signal-to-distortion ratio (SD-SDR) loss function at the output of the deep BMFMVDR filter. Instead of using the real and imaginary STFT coefficients as the input features as in [20], the STFT magnitude and the cosine of the STFT phase are used as the concatenated input features of the deep learning-based STCM estimators.

2.3. Hearing Loss Compensation

The hearing loss compensation stage is used for bilateral pure-tone audiogram-based compensation of hearing loss and further level adjustments. It consists of a spectral-domain multi-band dynamic range compressor (MBDRC) [23] implementing a noise gate, frequency- and hearing loss-dependent amplification and limitation of the maximum output level, and a volume control at the output. The STFT and filterbank parameters and the noise gate levels for the MBDRC were adopted from the CEC1 baseline system. The gains applied in the MBDRC were computed using the compressive *Camfit* gain prescription rule [28]. As an alternative to spectral-domain MBDRC, a simple broadband gain based on the half-gain rule (HGR) was considered, where the gain was computed as the pure-tone audiogram average at 500 Hz, 1000 Hz, and 2000 Hz divided by 2. The system also takes care of calibration and soft-clipping of the output signal, with settings adopted from the CEC1 baseline system.

3. Submitted Systems

Three systems were submitted to the challenge, where all systems were evaluated using objective metrics and two systems were evaluated using subjective listening experiments with hearing-impaired listeners (see Section 4). All systems use the wBLCMP beamformer (Section 2.1) as the first processing stage and hearing loss compensation (Section 2.3) as the last processing stage. The third submitted system additionally uses the deep BMFMVDR postfilter (Section 2.2) after the wBLCMP beamformer and before the hearing loss compensation stage.

- **E016**: Combination of wBLCMP beamformer and HGR-based hearing loss compensation.
- **E019**: Combination of wBLCMP beamformer and MBDRC.
- **E021**: Combination of wBLCMP beamformer, deep BMFMVDR postfilter and MBDRC.

3.1. Algorithm Settings

In this section, we briefly describe the settings of the used algorithms. Before processing, the microphone signals have been downsampled from 44.1 kHz to 16 kHz.

3.1.1. Weighted Binaural LCMP beamformer

The parameters of the STFT framework used for the wBLCMP beamformer are presented in Table 1. The filter length L_h of the wBLCMP beamformer was chosen to be 8 frames $\hat{=}\,20\,\mathrm{ms}$ with a delay τ of 2 frames $\hat{=}\,5\,\mathrm{ms}$ as a good compromise between performance and computational cost. The chosen shape parameter of $p = 0.5$ promotes sparsity less than $p = 0$, which slightly improved the performance. Since the RTFs of the target speaker and the interfering source are stationary we chose the smoothing constant $\gamma = 1$, corresponding to a growing window. The scaling factor $\beta_1 = 1$ corresponds to a distortionless constraint for the target speaker and $\beta_2 = 0.1$ was chosen to suppress the interfering source only partly to keep the spatial awareness of the acoustic scene.

Table 1: *Parameter values used in the wBLCMP beamformer.*

Parameter	Symbol	Value
STFT frame length		80 samples $\hat{=}\,5\mathrm{ms}$
STFT frame shift		40 samples $\hat{=}\,2.5\mathrm{ms}$
STFT window		sqrt-Hann

3.1.2. Deep Binaural Multi-Frame MVDR Filter

For the STFT framework we used the same parameters as for the wBLCMP beamformer. The deep BMFMVDR filter used a filter length of $N = 4$, and it was trained on the official CEC1 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. The choice of training on the official CEC1 training data instead of on the output of the wBLCMP beamformer was made based on preliminary experiments.

For the employed temporal convolutional networks (TCNs), we used 2 stacks of 8 layers each, with a kernel size of 3, resulting in a temporal receptive field of about $2.56\mathrm{s}$ and $3.02\mathrm{M}$ parameters. Note that the receptive field was deliberately chosen to be larger than 2s in order to cover the full interferer-only segment at the start of each CEC1 utterance.

3.1.3. Hearing Loss Compensation

For the hearing loss compensation stage, the parameters in Table 2 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.

Table 2: *Parameter values used in the hearing loss compensation stage for the submitted systems.*

	E016	E019	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

4. Results

In this section, we present the evaluation results provided by the CEC1 organizers based on the evaluation dataset. For each utterance in the evaluation dataset, a bilateral pure-tone audiogram detailing the hearing loss of a specific listener was provided, which was used in the hearing loss compensation stage to individualize the output signals for that specific listener.

For the objective evaluation, the CEC1 organizers processed

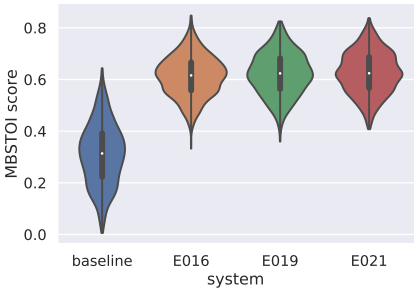

Figure 2: *Modified binaural short-term objective intelligibility results of the baseline system and the submitted systems on the evaluation dataset.*

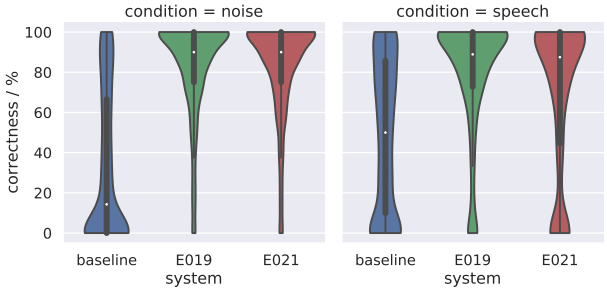

Figure 3: *Listening test results of the baseline system and the submitted systems chosen for the second evaluation stage on the evaluation dataset.*

the output signals of our submitted systems using a hearing loss model [29] that had previously been made available to the CEC1 participants, before estimating the speech intelligibility using the (also provided) modified binaural short-term objective intelligibility (MBSTOI) measure [30]. Figure 2 depicts a violin plot of 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, i.e., the submitted systems achieve an MBSTOI median score of approximately 0.62 while the baseline system achieves an MBSTOI median score of approximately 0.31. 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 deep BMFMVDR postfilter achieve a significant improvement in terms of speech intelligibility upon the system E016.

Based on these results and the algorithmic differences between the submitted systems, two of our submitted systems were allowed to proceed to the second evaluation stage. Aiming at evaluating the effect of the deep BMFMVDR postfilter, we selected systems E019 and E021, both including the wBLCMP beamformer and the MBDRC hearing loss compensation, where only E021 included the deep BMFMVDR postfilter. The second stage consisted of listening tests, in which hearing-impaired listeners were presented with the enhanced signals. For the *noise interferer condition*, the instructions were: "In the speech in noise test, you will hear a sentence and a loud distracting noise (e.g., a washing machine). You need to repeat what the talker is saying". For the *speech interferer condition*, the instructions were: "In the two talker test, you will hear two talkers speaking at the same time. One talker will start later than the other. You must repeat what this second talker is saying". Correctness was then evaluated by dividing the number of correctly identified words by the total number of words uttered by the target speaker.

The results of the perceptual evaluation stage are depicted as a violin plot in Figure 3 per interferer condition. First, it can be observed that both submitted systems greatly outperform the baseline system. More specifically, as stated in the official CEC1 results, the baseline system and our submitted systems E019 and E021 achieved a correctness score of approximately 33.2 %, 86.7 %, and 84.9 % in the noise interferer condition, and of approximately 51.2 %, 86.9 %, and 83.9 % in the speech interferer condition, respectively. Second, comparing the submitted systems, for the noise interferer condition the correctness scores are similar, whereas for the speech interferer condition there is a larger number of low correctness scores for system E021 than for system E019. A possible explanation of this observation is that, in some scenarios, system E021 reduced the first (interfering) speaker to such an extent that the listeners perceived the second (target) speaker as the first speaker, and thus did not repeat the understood words, as per the instructions quoted above. In such cases, keeping a small residual of the interfering signal may be preferred.

5. Conclusion

Aiming at improving speech intelligibility for hearing-impaired listeners in a reverberant scenario with a target speaker and an interfering source, in our contribution we proposed different combinations of a wBLCMP beamformer, a deep BMFMVDR postfilter, and a hearing loss compensation stage. The wBLCMP beamformer performs joint dereverberation and interferer reduction by minimizing a sparsity-promoting ℓ_p-norm cost function subject to linear constraints for both sources. These constraints aim at preserving the target speaker without distortion and reducing the interfering source in a controlled way, moreover preserving the binaural cues of both sources to preserve spatial awareness for the listener. To achieve additional interferer reduction, an optional binaural MFMVDR filter is used, where the required parameters are estimated using temporal convolutional networks by minimizing the SD-SDR loss function. The audiogram-based hearing loss compensation stage either uses a multi-band dynamic range compressor or a broadband gain based on the half-gain rule.

All submitted systems were shown to considerably improve speech intelligibility compared with the baseline system based on objective metrics and subjective listening experiments with hearing-impaired listeners. For the noise interferer condition, the best performing system yielded 86.7 % intelligibility, whereas for the speech interferer condition, the best performing system yielded 86.9 % intelligibility.

6. 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] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *The Journal of the Acoustical Society of America*, vol. 113, no. 6, p. 3233, 2003.

[3] A. Warzybok, J. Rennies, T. Brand, S. Doclo, and B. Kollmeier, "Effects of spatial and temporal integration of a single early reflection on speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. 269–282, Jan. 2013.

[4] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 331–342, Jul. 2006.

[5] M. Lavandier and J. F. Culling, "Speech segregation in rooms: Monaural, binaural, and interacting effects of reverberation on target and interferer," *The Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 2237–2248, Apr. 2008.

[6] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel Signal Enhancement Algorithms for Assisted Listening Devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[7] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[8] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

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

[10] 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.

[11] S. Markovich, S. Gannot, and I. Cohen, "Multichannel Eigenspace Beamforming in a Reverberant Noisy Environment With Multiple Interfering Speech Signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[12] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. Juang, "Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.

[13] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-Channel Linear Prediction-Based Speech Dereverberation With Sparse Priors," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1509–1520, Sep. 2015.

[14] ——, "Group sparsity for MIMO speech dereverberation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz NY, USA, Oct. 2015, pp. 1–5.

[15] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, S. Araki, T. Hori, and T. Nakatani, "Strategies for distant speech recognitionin reverberant environments," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–15, Jul. 2015.

[16] T. Nakatani and K. Kinoshita, "A Unified Convolutional Beamformer for Simultaneous Denoising and Dereverberation," *IEEE Signal Processing Letters*, vol. 26, no. 6, pp. 903–907, Jun. 2019.

[17] C. Boeddeker, T. Nakatani, K. Kinoshita, and R. Haeb-Umbach, "Jointly Optimal Dereverberation and Beamforming," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain, May 2020, pp. 216–220.

[18] H. Gode, M. Tammen, and S. Doclo, "Joint multi-channel dereverberation and noise reduction using a unified convolutional beamformer with sparse priors," in *Proc. ITG Conference on Speech Communication*, Kiel, Germany, Sep. 2021.

[19] 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: IEEE, Sep. 2020, pp. 1–6.

[20] 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 (ICASSP)*, Toronto, Canada, June 2021, pp. 8443–8447.

[21] 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.

[22] S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," *arXiv:1803.01271 [cs]*, Mar. 2018.

[23] G. Grimm, T. Herzke, S. D. Ewert, and V. Hohmann, "Implementation and Evaluation of an Experimental Hearing Aid Dynamic Range Compressor," in *Proc. German Annual Conference on Acoustics (DAGA)*, Nuremberg, Germany, 2015, p. 4.

[24] Y. Avargel and I. Cohen, "On Multiplicative Transfer Function Approximation in the Short-Time Fourier Transform Domain," *IEEE Signal Processing Letters*, vol. 14, no. 5, pp. 337–340, May 2007.

[25] B. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Trans. on Signal Processing*, vol. 47, no. 1, pp. 187–200, Jan. 1999.

[26] S. Markovich-Golan, S. Gannot, and W. Kellermann, "Performance analysis of the covariance-whitening and the covariance-subtraction methods for estimating the relative transfer function," in *Proc. European Signal Processing Conference*, Rome, Italy, Sep. 2018, pp. 2499–2503.

[27] Y. A. Huang and J. Benesty, "A Multi-Frame Approach to the Frequency-Domain Single-Channel Noise Reduction Problem," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1256–1269, May 2012.

[28] 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.

[29] Y. Nejime and B. C. Moore, "Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise," *The Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 603–615, Jul. 1997.

[30] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Communication*, vol. 102, pp. 1–13, Sep. 2018.