Imposing Correlation Structures for Deep Binaural Spatio-Temporal Wiener Filtering

Marvin Tammen, *Graduate Student Member, IEEE*, and Simon Doclo, *Senior Member, IEEE*

***Abstract*—To improve speech quality and intelligibility in environments with noise and interfering sounds, binaural speech enhancement algorithms use the microphone signals from both the left and the right hearing device to generate an enhanced output signal for each ear. As a multi-frame extension of the binaural multi-channel Wiener filter, in this paper we consider the binaural spatio-temporal Wiener filter (STWF) in the short-time Fourier transform domain, which requires estimates of the highly time-varying spatio-temporal correlations of the speech and interference components. To this end, the binaural STWF is embedded into an end-to-end supervised learning framework, where temporal convolutional networks estimate the required quantities, i.e., the inverse spatio-temporal correlation matrices of the interference component and the spatio-temporal correlation vectors and power spectral densities of the speech components. In this paper, we impose spatio-temporal correlation structure on these quantities and relate them between the left and the right hearing device, aiming to reduce computational complexity while maintaining speech enhancement and interaural cue preservation performance. Assuming that the spatial correlation of the speech component is stationary over a small number of frames, we propose to decompose the spatio-temporal correlation vectors as the Kronecker product of a relative transfer function vector and a temporal correlation vector, either considering a global reference microphone or a reference microphone for each hearing device. In addition, we consider a deep bilateral STWF by neglecting the spatio-temporal correlations of the speech and interference components between both devices. The imposed spatio-temporal correlation structures greatly differ in the number of parameters that need to be estimated. The performance of causal versions of the deep binaural and bilateral STWF algorithms is evaluated based on both simulated and measured binaural room impulse responses (BRIRs) as well as diverse speech and noise sources. The simulation results demonstrate that the proposed spatio-temporal correlation structures significantly reduce the computational complexity of the binaural STWF while yielding a similar speech enhancement and interaural cue preservation performance compared to not imposing any spatio-temporal correlation structure. Furthermore, the results confirm that the deep binaural STWF outperforms the binaural Conv-TasNet algorithm as well as an algorithm that directly estimates the binaural multi-frame filter coefficients, while approaching the performance of the non-causal binaural complex convolutional transformer network (BCCTN) algorithm.**

***Index Terms*—Multi-channel wiener filter, multi-frame filtering, spatio-temporal correlation structure, supervised learning, binaural speech enhancement.**

I. Introduction

In many everyday speech communication scenarios, we are confronted with undesired noise and interfering sounds, reducing speech quality and speech intelligibility. Hence, algorithms for hearing devices are required to extract the target speaker and reduce noise and interference from the recorded microphone signals. In principle, using hearing devices on both ears can generate an important advantage, both from a signal processing perspective, since all microphone signals from both devices can be used, as well as from a perceptual perspective, since interaural cues can be exploited by the auditory system [1], [2]. An important distinction exists between *bilateral* systems, where both devices operate independently, and *binaural* systems, where microphone signals from both devices are processed and combined in each device. Several binaural speech enhancement algorithms have been proposed, which can be broadly categorized into statistical model-based approaches (e.g., [2], [3], [4], [5], [6], [7]) and supervised learning-based approaches (e.g., [8], [9], [10], [11], [12], [13], [14], [15]). Most approaches estimate a clean speech signal for the left and the right hearing device by applying a mask or filter to the noisy microphone signals in a transform domain, e.g., the short-time Fourier transform (STFT) domain or a learned transform domain. In general, supervised learning-based approaches tend to outperform statistical model-based approaches, especially in reducing non-stationary interference.

In this paper, we will focus on supervised learning-based algorithms in the STFT domain. Aiming at exploiting temporal correlations across successive STFT frames, multi-frame algorithms for both single- and multi-microphone speech enhancement have recently gained popularity [15], [16], [17], [18], [19], [20], [21], [22], [23]. Instead of directly estimating the multi-frame filter coefficients using a deep neural network (DNN) as in [16], [17], several procedures have been proposed to impose a certain structure on the multi-frame filter. For single-microphone speech enhancement, a frequently used filter structure is the multi-frame minimum variance distortionless response (MVDR) filter [23], [24], [25]. In each STFT bin, the multi-frame MVDR filter estimates the target speech STFT coefficient by minimizing

Received 15 September 2024; revised 20 January 2025; accepted 23 February 2025. Date of publication 5 March 2025; date of current version 21 March 2025. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through Germany's Excellence Strategy - EXC 2177/1 - under Grant 390895286 and Grant 352015383 - SFB 1330 B2. The associate editor coordinating the review of this article and approving it for publication was Prof. Jong Won Shin. *(Corresponding author: Marvin Tammen.)*

The authors are with the Department of Medical Physics and Acoustics and the Cluster of Excellence Hearing4all, Carl von Ossietzky Universität Oldenburg, 26129 Oldenburg, Germany (e-mail: marvin.tammen@uni-oldenburg.de; simon.doclo@uni-oldenburg.de).

Digital Object Identifier 10.1109/TASLPRO.2025.3548454

the output interference power spectral density (PSD) while preserving the temporal correlation of the speech component, thereby requiring estimates of the highly time-varying speech temporal correlation vector and the inverse interference temporal correlation matrix.[1] For multi-microphone speech enhancement, frequently used filter structures include the spatio-temporal MVDR filter and the STWF [20], [22]. These spatio-temporal filters require estimates of spatio-temporal covariance matrices (STCMs) and the speech spatio-temporal correlation vector (STCV). On the one hand, estimation approaches have been proposed where the spatio-temporal MVDR filter or the STWF are decoupled from the training of the DNN. For example, in [22] a DNN first estimates the target speech component at a reference microphone, which is subsequently used to compute the STWF. On the other hand, approaches have been proposed where the spatio-temporal MVDR filter or the STWF are fully integrated into the training of the DNN [15], [20], [21]. For example, in [20] a DNN estimates the speech STCV and the inverse interference STCM required by the spatio-temporal MVDR filter. In [15], [23], [26] it was shown that imposing a Hermitian positive-definite structure on the (inverse) STCMs, e.g., using a Cholesky decomposition, improves the performance in terms of objective speech enhancement metrics. This was shown for the speech and interference spatial covariance matrices in the single-frame multi-microphone MVDR filter [26], the noisy and interference temporal covariance matrices in the multi-frame single-microphone MVDR filter [23], as well as the inverse interference STCM in a binaural extension of the spatio-temporal MVDR filter [15]. It should be noted that none of the aforementioned approaches imposed a specific structure on the speech STCV.

In this paper, we focus on a causal binaural extension of the STWF, which estimates the target speech component in a reference microphone at the left and the right hearing device by filtering all available microphone signals from both devices. To reduce computational complexity while maintaining speech enhancement and interaural cue preservation performance, we propose to impose various spatio-temporal correlation structures on the required quantities during the estimation process. Specifically, we consider the decomposition of the binaural STWF into a binaural spatio-temporal MVDR filter and a spectral postfilter, hence requiring estimates of the speech STCVs, the speech PSDs, and the inverse interference STCMs. Please note that in principle these quantities need to be estimated both for the filter estimating the target speech component at the left device as well as for the filter estimating the target speech component at the right device. The binaural STWF is embedded into an end-to-end supervised learning framework, where all quantities are estimated using temporal convolutional networks (TCNs). In addition to imposing a Hermitian positive-definite structure on the inverse interference STCMs, the main objective of this paper is to investigate the potential of imposing spatio-temporal correlation structure on the speech STCVs and the inverse interference STCMs. We propose several procedures which mainly differ in terms of the relation between the microphones, particularly between the left and the right hearing device, and the number of parameters that need to be estimated. First, assuming that the spatial correlation of the speech component is stationary over the length of the multi-frame filter, the speech STCVs can be decomposed as the Kronecker product of a relative transfer function vector and a temporal correlation vector. We either consider a single "global" reference microphone, requiring the speech temporal correlation vector to be estimated only for this microphone, or a reference microphone for each hearing device, requiring (left and right) speech temporal correlation vectors to be estimated for both reference microphones. The STCV structure considering two reference microphones involves more parameters than the STCV structure considering a single reference microphone, but it allows for more degrees of freedom. Second, we propose to replace the left and right interference STCMs by a common interference STCM, as the difference between both STCMs can be assumed to be negligible. In addition, we consider a bilateral STWF by assuming no spatio-temporal correlation between both hearing devices, both for the speech STCVs and for the interference STCM. To train and evaluate the deep bilateral STWF and the deep binaural STWF using the proposed spatio-temporal correlation structures, we constructed matched datasets using diverse speech and noise sources from the DNS 1 and DNS 2 challenges [27], [28] as well as simulated binaural room impulse responses from the Clarity Enhancement Challenge (CEC) 1 [29]. In addition, to evaluate the generalization capabilities of the considered algorithms, we considered a mismatched evaluation dataset from CEC 3 that comprises noise backgrounds and room impulse responses (RIRs) recorded in complex environments as well as simulated head rotation. Simulation results show that the binaural STWF using a combination of the speech STCV structure considering two reference microphones and a common interference STCM significantly reduces the computational complexity while yielding a similar speech enhancement and interaural cue preservation performance compared to not imposing any spatio-temporal correlation structure. Furthermore, simulation results demonstrate that this deep binaural STWF outperforms the deep bilateral STWF as well as two state-of-the-art binaural speech enhancement algorithms, namely the deep filtering algorithm [16] (which directly estimates the binaural multi-frame filter coefficients) and the binaural Conv-TasNet algorithm [9], while approaching the performance of the non-causal BCCTN algorithm [14].

The remainder of the paper is organized as follows. In Section II, we describe the signal model and introduce the binaural STWF. In Section III, we propose several spatio-temporal correlation structures for the speech STCVs and the inverse interference STCMs. In Section IV, we describe the supervised learning-based approach to estimate the required quantities, taking into account these correlation structures. The simulation setup, the approach to validate the proposed correlation structures, and the simulation results for the proposed deep binaural and bilateral STWFs as well as for several baseline algorithms are presented and discussed in Sections V, VI, and VII.

[1] In multi-frame MVDR filtering, the interference component is typically defined to include both noise components and speech components that are uncorrelated with the current target speech STFT coefficient [24], [25].

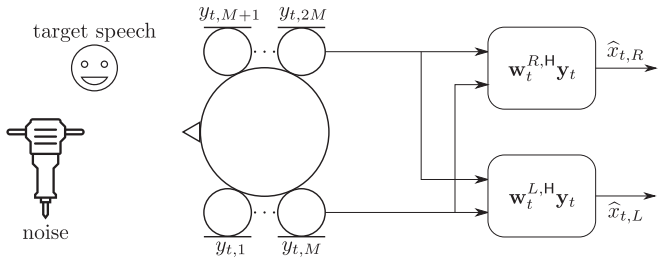


Fig. 1. Acoustic scenario with a target speech source and noise as well as binaural processing scheme, estimating the target speech component at the left and the right hearing device by filtering all available microphone signals.

II. Spatio-Temporal Wiener Filter for Binaural Speech Enhancement

A. Signal Model

We consider an acoustic scenario with a single speech source and background noise in a reverberant room, recorded by binaural hearing devices with M microphones each, i.e., a total of $2M$ microphones (see Fig. 1). We assume that all microphone signals are synchronized and transmitted (e.g., via a wireless link) between the hearing devices without transmission delay and quantization errors. In the STFT domain, the m-th noisy microphone signal $y_{f,t,m}$ at frequency bin f and time frame t is given by

$$y_{f,t,m} = x_{f,t,m} + n_{f,t,m}, \quad m \in \{1, \dots, 2M\}, \tag{1}$$

where $x_{f,t,m}$ and $n_{f,t,m}$ denote the (reverberant) target speech component and the noise component, respectively. Multi-frame speech enhancement algorithms [15], [16], [17], [18], [19], [20], [21], [22], [23] use the noisy multi-frame vector $\mathbf{y}_{f,t}$, where N denotes the multi-frame filter length, defined as

$$\mathbf{y}_{f,t} = \begin{bmatrix} \bar{\mathbf{y}}_{f,t,1}^{\mathsf{T}} & \dots & \bar{\mathbf{y}}_{f,t,2M}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} \in \mathbb{C}^{2MN}, \tag{2}$$

with $\cdot^{\mathsf{T}}$ denoting the transpose operator and

$$\bar{\mathbf{y}}_{f,t,m} = \begin{bmatrix} y_{f,t,m} & \dots & y_{f,t-N+1,m} \end{bmatrix}^{\mathsf{T}}. \tag{3}$$

Since each frequency bin is processed independently, the index f will be omitted in the remainder of this paper. Using (1), the multi-frame vector in (2) can be written as

$$\mathbf{y}_t = \mathbf{x}_t + \mathbf{n}_t, \tag{4}$$

where $\mathbf{x}_t$ and $\mathbf{n}_t$ are defined similarly as in (2). The target speech components $\widehat{x}_{t,L}$ and $\widehat{x}_{t,R}$ at both hearing devices are defined (without loss of generality) at reference microphones $L = 1$ and $R = M + 1$ (see Fig. 1). The target speech components are estimated by processing all available microphone signals with the binaural spatio-temporal filters $\mathbf{w}_t^L \in \mathbb{C}^{2MN}$ (for the left device) and $\mathbf{w}_t^R \in \mathbb{C}^{2MN}$ (for the right device), i.e.,

$$\widehat{x}_{t,L} = \mathbf{w}_t^{L,\mathsf{H}}\mathbf{y}_t, \quad \widehat{x}_{t,R} = \mathbf{w}_t^{R,\mathsf{H}}\mathbf{y}_t, \tag{5}$$

where the binaural spatio-temporal filters are defined similarly as in (2), and $\cdot^{\mathsf{H}}$ denotes the complex conjugate transpose operator.

Assuming that the speech and noise components are spatio-temporally uncorrelated, the noisy STCM $\mathbf{\Phi}_{y,t} = \mathcal{E}\{\mathbf{y}_t\mathbf{y}_t^{\mathsf{H}}\} \in \mathbb{C}^{2MN\times 2MN}$, with $\mathcal{E}\{\cdot\}$ the expectation operator, can be written as

$$\mathbf{\Phi}_{y,t} = \mathbf{\Phi}_{x,t} + \mathbf{\Phi}_{n,t}, \tag{6}$$

where $\mathbf{\Phi}_{x,t} = \mathcal{E}\{\mathbf{x}_t\mathbf{x}_t^{\mathsf{H}}\}$ and $\mathbf{\Phi}_{n,t} = \mathcal{E}\{\mathbf{n}_t\mathbf{n}_t^{\mathsf{H}}\}$. The matrix $\mathbf{\Phi}_{y,t}$ can be partitioned as

$$\mathbf{\Phi}_{y,t} = \begin{bmatrix} \tilde{\mathbf{\Phi}}_{y,t}^{LL} & \tilde{\mathbf{\Phi}}_{y,t}^{LR} \\ \tilde{\mathbf{\Phi}}_{y,t}^{LR,\mathsf{H}} & \tilde{\mathbf{\Phi}}_{y,t}^{RR} \end{bmatrix}, \tag{7}$$

where the submatrix $\tilde{\mathbf{\Phi}}_{y,t}^{LL}$ contains only spatio-temporal correlations of the microphones at the left hearing device, the submatrix $\tilde{\mathbf{\Phi}}_{y,t}^{RR}$ contains only spatio-temporal correlations of the microphones at the right hearing device, and the submatrix $\tilde{\mathbf{\Phi}}_{y,t}^{LR}$ contains spatio-temporal correlations between the contralateral microphones.

In order to exploit speech correlations across successive time frames, it was proposed in [24], [25] to decompose the (single-microphone) multi-frame speech vector into a temporally correlated component and a temporally uncorrelated component. Similarly, the (multi-microphone) multi-frame speech vector $\mathbf{x}_t$ can be decomposed into a spatio-temporally correlated component and a spatio-temporally uncorrelated component $\mathbf{x}'_{t,m}$ w.r.t. the speech STFT coefficient $x_{t,m}$ at the m-th microphone, i.e.,

$$\mathbf{x}_t = \underbrace{\boldsymbol{\gamma}_t^m x_{t,m}}_{\text{correlated}} + \underbrace{\mathbf{x}'_{t,m}}_{\text{uncorrelated}}, \tag{8}$$

where the speech STCV $\boldsymbol{\gamma}_t^m$ describes the correlation between the N most recent speech STFT coefficients at each microphone and the current speech STFT coefficient at the m-th microphone. The speech STCV is defined as

$$\boxed{\boldsymbol{\gamma}_t^m = \frac{\mathcal{E}\left\{\mathbf{x}_t x_{t,m}^*\right\}}{\mathcal{E}\left\{|x_{t,m}|^2\right\}} = \frac{\mathbf{\Phi}_{x,t}\mathbf{e}_m}{\phi_t^m} \in \mathbb{C}^{2MN}} \tag{9}$$

where $\cdot^*$ denotes the complex conjugate operator, $\mathbf{e}_m$ denotes a selection vector with the element corresponding to the current frame at the m-th microphone equal to 1 and the other elements equal to 0, and $\phi_t^m = \mathbf{e}_m^{\mathsf{T}}\mathbf{\Phi}_{x,t}\mathbf{e}_m$ denotes the speech PSD at the m-th microphone. It can be easily shown that the element of the speech STCV corresponding to the current frame at the m-th microphone is equal to 1, i.e.,

$$\mathbf{e}_m^{\mathsf{T}}\boldsymbol{\gamma}_t^m = 1. \tag{10}$$

It should be noted that the speech STCV depends both on the temporal correlation of the speech component, which is highly time-varying, and on the spatial correlation, which can be assumed to be less time-varying than the temporal correlation. The decomposition in (8) can be carried out for both reference microphones, i.e.,

$$\boxed{\mathbf{x}_t = \boldsymbol{\gamma}_t^L x_{t,L} + \mathbf{x}'_{t,L} = \boldsymbol{\gamma}_t^R x_{t,R} + \mathbf{x}'_{t,R}} \tag{11}$$

where $\boldsymbol{\gamma}_t^L$ denotes the left speech STCV and $\boldsymbol{\gamma}_t^R$ denotes the right speech STCV.

Substituting (11) into (4), the noisy multi-frame vector can be written as

$$\mathbf{y}_t = \boldsymbol{\gamma}_t^{\nu} x_{t,\nu} + \underbrace{\mathbf{x}'_{t,\nu} + \mathbf{n}_t}_{=:\mathbf{i}_{t,\nu}}, \quad \nu \in \{L, R\}, \tag{12}$$

where $\mathbf{i}_{t,\nu}$ denotes the interference vector, containing both the uncorrelated speech component $\mathbf{x}'_{t,\nu}$ and the noise component $\mathbf{n}_t$. Using (12), the noisy STCM in (6) can be written as

$$\boxed{\boldsymbol{\Phi}_{y,t} = \phi_t^{\nu} \boldsymbol{\gamma}_t^{\nu} \boldsymbol{\gamma}_t^{\nu,\mathsf{H}} + \underbrace{\boldsymbol{\Phi}_{x',t}^{\nu} + \boldsymbol{\Phi}_{n,t}}_{=\boldsymbol{\Phi}_{i,t}^{\nu}}, \quad \nu \in \{L, R\}} \tag{13}$$

where $\boldsymbol{\Phi}_{x',t}^{\nu} = \mathcal{E}\{\mathbf{x}'_{t,\nu}\mathbf{x}'^{\mathsf{H}}_{t,\nu}\}$ and $\boldsymbol{\Phi}_{i,t}^{\nu} = \mathcal{E}\{\mathbf{i}_{t,\nu}\mathbf{i}_{t,\nu}^{\mathsf{H}}\}$ denote the uncorrelated speech STCM and the interference STCM, respectively. It should be noted that, in general, the speech PSDs ϕ_t^{ν}, the speech STCVs $\boldsymbol{\gamma}_t^{\nu}$, the uncorrelated speech STCMs $\boldsymbol{\Phi}_{x',t}^{\nu}$, and the interference STCMs $\boldsymbol{\Phi}_{i,t}^{\nu}$ are different for the left hearing device ($\nu = L$) and the right hearing device ($\nu = R$).

B. Binaural Spatio-Temporal Wiener Filter

In [4], [30] the binaural multi-channel Wiener filter was proposed, which aims at minimizing the mean square error between the binaural output signals and the target speech components at both reference microphones, only considering spatial correlations. In this paper, we consider a multi-frame extension of the binaural multi-channel Wiener filter, termed binaural STWF, which considers both spatial as well as temporal correlations. The binaural STWF can also be considered as a binaural extension of the STWF proposed in [24]. The binaural spatio-temporal filters $\mathbf{w}_t^L$ and $\mathbf{w}_t^R$ are computed by minimizing the cost function

$$J(\mathbf{w}_t^L, \mathbf{w}_t^R) = \mathcal{E}\left\{ \left\| \begin{matrix} x_{t,L} - \mathbf{w}_t^{L,\mathsf{H}} \mathbf{y}_t \\ x_{t,R} - \mathbf{w}_t^{R,\mathsf{H}} \mathbf{y}_t \end{matrix} \right\|_2^2 \right\}, \tag{14}$$

yielding

$$\mathbf{w}_t^{\nu} = \boldsymbol{\Phi}_{y,t}^{-1} \boldsymbol{\Phi}_{x,t} \mathbf{e}_{\nu}, \quad \nu \in \{L, R\}. \tag{15}$$

Using (13) and the fact that $\boldsymbol{\Phi}_{x',t}^{\nu} \mathbf{e}_{\nu} = \mathbf{0}$, it can be easily shown that both STWF vectors in (15) can be decomposed as a spatio-temporal MVDR filter [20] and a real-valued scalar postfilter, i.e.,

$$\boxed{\mathbf{w}_t^{\nu} = \underbrace{\frac{\left(\boldsymbol{\Phi}_{i,t}^{\nu}\right)^{-1} \boldsymbol{\gamma}_t^{\nu}}{\boldsymbol{\gamma}_t^{\nu,\mathsf{H}} \left(\boldsymbol{\Phi}_{i,t}^{\nu}\right)^{-1} \boldsymbol{\gamma}_t^{\nu}}}_{\text{spatio-temporal MVDR}} \underbrace{\frac{\phi_t^{\nu}}{\phi_t^{\nu} + \boldsymbol{\gamma}_t^{\nu,\mathsf{H}} \left(\boldsymbol{\Phi}_{i,t}^{\nu}\right)^{-1} \boldsymbol{\gamma}_t^{\nu}}}_{\text{postfilter}}} \tag{16}$$

with $\nu \in \{L, R\}$. The spatio-temporal MVDR filter minimizes the interference power while preserving the spatio-temporal correlation of the speech component, with the postfilter providing additional noise reduction at the cost of allowing for some speech distortion. The performance of the binaural STWF strongly depends on how well the required quantities, i.e., the left and right inverse interference STCMs $(\boldsymbol{\Phi}_{i,t}^{L})^{-1}$ and $(\boldsymbol{\Phi}_{i,t}^{R})^{-1}$, the left and right speech STCVs $\boldsymbol{\gamma}_t^{L}$ and $\boldsymbol{\gamma}_t^{R}$, as well as the left and right speech PSDs ϕ_t^{L} and ϕ_t^{R} are estimated from the noisy STFT coefficients. Similarly as in [15], [20], [23], [26], in this paper we embed the binaural STWF in a supervised learning framework, estimating the required quantities with DNNs (see Section IV). In addition, we investigate imposing different spatio-temporal structures on the speech STCVs and the interference STCMs, as described in the following section.

III. Spatio-Temporal Correlation Structures

The quantities required by the binaural STWF are determined by both temporal and spatial correlations. On the one hand, the temporal correlations of the speech and interference components can vary drastically across a small number of time frames. On the other hand, the spatial correlations of the speech and interference components mainly depend on the acoustic scene, i.e., the positions of the listener and the speech and noise sources, which can be assumed to be stationary across a small number of time frames. In this section, we propose several spatio-temporal structures for the speech STCVs and the interference STCMs, relating these quantities between the left and the right hearing device. First, assuming spatial stationarity of the speech component over a small number of time frames, in Section III-A we impose spatial structure on the speech STCVs. Second, assuming that the uncorrelated speech components are negligible, in Section III-B we set the left and the right interference STCM equal to each other. Third, in Section III-C we assume no correlation between the left and right hearing devices for both the speech and interference components. The considered structures greatly differ in the number of parameters that need to be estimated. As will be demonstrated by the simulation results in Section VII, imposing structure on the speech STCVs and the interference STCMs is beneficial in terms of computational complexity.

A. Speech Correlation Vectors

In many multi-microphone speech enhancement algorithms, multiplicative relative transfer functions (RTFs) have been utilized to model the relationship between the speech STFT coefficients at all microphones [3]. Assuming the microphone with index ν as the reference microphone, the (single-frame) multi-microphone speech vector can be written as

$$\begin{bmatrix} x_{t,1} \\ x_{t,2} \\ \vdots \\ x_{t,2M} \end{bmatrix} = \begin{bmatrix} h_{t,1}^{\nu} \\ h_{t,2}^{\nu} \\ \vdots \\ h_{t,2M}^{\nu} \end{bmatrix} x_{t,\nu}, \tag{17}$$

with $h_{t,m}^{\nu}$ denoting the RTF between the m-th microphone and the reference microphone, defined as

$$h_{t,m}^{\nu} = \frac{\mathcal{E}\{x_{t,m} x_{t,\nu}^*\}}{\mathcal{E}\{|x_{t,\nu}|^2\}} = \frac{\mathbf{e}_m^{\mathsf{T}} \boldsymbol{\Phi}_{x,t} \mathbf{e}_{\nu}}{\phi_t^{\nu}}, \tag{18}$$

such that $h_{t,\nu}^{\nu} = 1$. Similarly, the multi-frame speech vector at the m-th microphone can be written as

$$\bar{\mathbf{x}}_{t,m} = \begin{bmatrix} x_{t,m} \\ x_{t-1,m} \\ \vdots \\ x_{t-N+1,m} \end{bmatrix} = \begin{bmatrix} h_{t,m}^{\nu} x_{t,\nu} \\ h_{t-1,m}^{\nu} x_{t-1,\nu} \\ \vdots \\ h_{t-N+1,m}^{\nu} x_{t-N+1,\nu} \end{bmatrix}. \tag{19}$$

Assuming that the RTFs are constant over N frames (i.e., $h^{\nu}_{t,m}, h^{\nu}_{t-1,m}, \ldots, h^{\nu}_{t-N+1,m}$ are equal), the multi-frame speech vector in (19) can be written as [31]

$$\bar{\mathbf{x}}_{t,m} = h^{\nu}_{t,m}\bar{\mathbf{x}}_{t,\nu}. \tag{20}$$

In the following, we will either consider a single "global" reference microphone for both hearing devices or a reference microphone for each hearing device.

1) Global Relative Transfer Function: Without loss of generality, we choose the reference microphone on the left hearing device (with index $L = 1$) as the global reference microphone. Using (20), the multi-microphone multi-frame speech vector in (4) can then be modeled as

$$\mathbf{x}_t = \begin{bmatrix} \bar{\mathbf{x}}_{t,1} \\ \bar{\mathbf{x}}_{t,2} \\ \vdots \\ \bar{\mathbf{x}}_{t,2M} \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{x}}_{t,1} \\ h^1_{t,2}\bar{\mathbf{x}}_{t,1} \\ \vdots \\ h^1_{t,2M}\bar{\mathbf{x}}_{t,1} \end{bmatrix} = \mathbf{h}_{t,\text{glob}} \otimes \bar{\mathbf{x}}_{t,1}, \tag{21}$$

where the global RTF vector $\mathbf{h}_{t,\text{glob}}$ contains the RTFs between all microphones and the global reference microphone, i.e.,

$$\mathbf{h}_{t,\text{glob}} = \begin{bmatrix} 1 & h^1_{t,2} & \ldots & h^1_{t,2M} \end{bmatrix}^{\mathsf{T}} \in \mathbb{C}^{2M}, \tag{22}$$

and $\otimes$ denotes the Kronecker product. Using (21), the left and right speech STCVs in (9) can be written as

$$\boldsymbol{\gamma}^{\nu}_{t,\text{glob}} = \frac{\mathcal{E}\left\{\mathbf{h}_{t,\text{glob}} \otimes \bar{\mathbf{x}}_{t,1}x^*_{t,\nu}\right\}}{\phi^{\nu}_t} = \mathbf{h}_{t,\text{glob}} \otimes \bar{\boldsymbol{\gamma}}^{\nu}_{t,1}, \tag{23}$$

with $\nu \in \{L, R\}$, and where the speech temporal correlation vector $\bar{\boldsymbol{\gamma}}^{\nu}_{t,1}$ describes the correlation between the N most recent speech STFT coefficients at the global reference microphone (with index $L = 1$) and the current target speech STFT coefficient ($x_{t,L}$ or $x_{t,R}$), i.e.,

$$\bar{\boldsymbol{\gamma}}^{\nu}_{t,1} = \frac{\mathcal{E}\left\{\bar{\mathbf{x}}_{t,1}x^*_{t,\nu}\right\}}{\phi^{\nu}_t} \in \mathbb{C}^N, \quad \nu \in \{L, R\}. \tag{24}$$

When imposing the global RTF structure on the speech STCVs in (23), the speech STCVs can be interpreted as being decomposed into a spatial factor ($\mathbf{h}_{t,\text{glob}}$) and a temporal factor ($\bar{\boldsymbol{\gamma}}^{\nu}_{t,1}$). Furthermore, for two of the quantities to be estimated, the global RTF structure yields an explicit relation between the left and the right hearing device: First, since the reference microphones on both hearing devices are related as $x_{t,R} = h^1_{t,R}x_{t,L}$, the left and right speech PSDs are related as

$$\phi^R_{t,\text{glob}} = |h^1_{t,R}|^2\phi^L_{t,\text{glob}}, \tag{25}$$

where the global RTF $h^1_{t,R}$ is an element of $\mathbf{h}_{t,\text{glob}}$ in (22). Second, the left and right speech STCVs in (23) are related as

$$\boldsymbol{\gamma}^R_{t,\text{glob}} = \frac{\mathcal{E}\left\{\mathbf{x}_t x^*_{t,R}\right\}}{\mathcal{E}\left\{|x_{t,R}|^2\right\}} = \frac{1}{h^1_{t,R}}\frac{\mathcal{E}\left\{\mathbf{x}_t x^*_{t,L}\right\}}{\mathcal{E}\left\{|x_{t,L}|^2\right\}} = \frac{1}{h^1_{t,R}}\boldsymbol{\gamma}^L_{t,\text{glob}}, \tag{26}$$

i.e., the left and right speech STCVs are related by a complex-valued scalar (so they are parallel). It should be realized that—although both vectors may differ in amplitude and phase—the relation between microphones is the same for both hearing devices. The global RTF structure is visualized in Fig. 2(b), assuming $M = 2$ microphones per hearing device.

The model in (21) assumes fully correlated speech components between the global reference microphone and all microphones of both hearing devices, which is a common assumption in binaural speech enhancement algorithms [3], [4]. However, depending on the STFT frame length, this assumption may be violated in practice due to large inter-microphone distances and reverberation, especially when considering the correlation between the global reference microphone and the contralateral microphones. This motivates the investigation of an alternative structure in the next section.

2) Ipsilateral Relative Transfer Function: As a less restrictive alternative to the global RTF structure in (21), we propose to use this model for each hearing device independently, i.e., using the microphone with index L as the reference for the microphones of the left hearing device and the microphone with index R as the reference for the microphones of the right hearing device. The multi-microphone multi-frame speech vector can then be modeled as

$$\mathbf{x}_t = \begin{bmatrix} \mathbf{h}^L_{t,\text{ipsi}} \otimes \bar{\mathbf{x}}_{t,L} \\ \mathbf{h}^R_{t,\text{ipsi}} \otimes \bar{\mathbf{x}}_{t,R} \end{bmatrix}, \tag{27}$$

where the ipsilateral RTF vectors, defined as

$$\mathbf{h}^L_{t,\text{ipsi}} = \begin{bmatrix} 1 & h^L_{t,2} & \ldots & h^L_{t,M} \end{bmatrix}^{\mathsf{T}} \in \mathbb{C}^M \tag{28}$$

$$\mathbf{h}^R_{t,\text{ipsi}} = \begin{bmatrix} 1 & h^R_{t,M+2} & \ldots & h^R_{t,2M} \end{bmatrix}^{\mathsf{T}} \in \mathbb{C}^M, \tag{29}$$

relate the speech component at the microphones of the left hearing device to the left reference microphone and the speech component at the microphones of the right hearing device to the right reference microphone. Using (27) in (9), the left and right speech STCVs can be written as

$$\boldsymbol{\gamma}^{\nu}_{t,\text{ipsi}} = \frac{1}{\phi^{\nu}_t}\mathcal{E}\left\{\begin{bmatrix} \mathbf{h}^L_{t,\text{ipsi}} \otimes \bar{\mathbf{x}}_{t,L} \\ \mathbf{h}^R_{t,\text{ipsi}} \otimes \bar{\mathbf{x}}_{t,R} \end{bmatrix} x^*_{t,\nu}\right\} = \begin{bmatrix} \mathbf{h}^L_{t,\text{ipsi}} \otimes \bar{\boldsymbol{\gamma}}^{\nu}_{t,L} \\ \mathbf{h}^R_{t,\text{ipsi}} \otimes \bar{\boldsymbol{\gamma}}^{\nu}_{t,R} \end{bmatrix}, \tag{30}$$

with $\nu \in \{L, R\}$, and where the speech temporal correlation vectors $\bar{\boldsymbol{\gamma}}^{\nu}_{t,L}$ and $\bar{\boldsymbol{\gamma}}^{\nu}_{t,R}$ are defined similarly as in (24), i.e.,

$$\bar{\boldsymbol{\gamma}}^{\nu}_{t,L} = \frac{\mathcal{E}\left\{\bar{\mathbf{x}}_{t,L}x^*_{t,\nu}\right\}}{\phi^{\nu}_t}, \quad \bar{\boldsymbol{\gamma}}^{\nu}_{t,R} = \frac{\mathcal{E}\left\{\bar{\mathbf{x}}_{t,R}x^*_{t,\nu}\right\}}{\phi^{\nu}_t}. \tag{31}$$

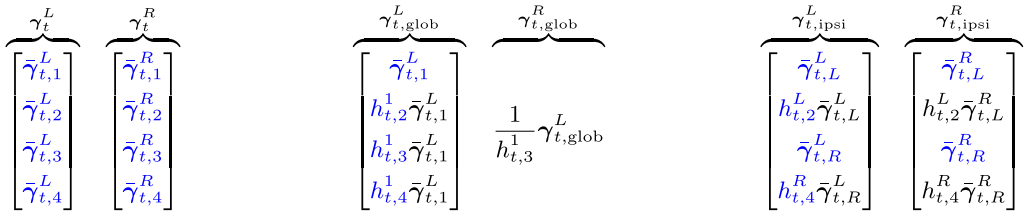
It should be noted that the ipsilateral RTF structure comprises four speech temporal correlation vectors, whereas the global RTF structure only comprises one speech temporal correlation vector in (24). The ipsilateral RTF structure is visualized in Fig. 2(c).

B. Interference Covariance Matrices

The left and right interference STCMs are defined as

$$\boldsymbol{\Phi}^{\nu}_{i,t} = \boldsymbol{\Phi}^{\nu}_{x',t} + \boldsymbol{\Phi}_{n,t}, \quad \nu \in \{L, R\}. \tag{32}$$

Assuming the uncorrelated speech STCMs $\boldsymbol{\Phi}^{\nu}_{x',t}$ to be negligible compared to the noise STCM $\boldsymbol{\Phi}_{n,t}$, the left and right interference STCMs $\boldsymbol{\Phi}^L_{i,t}$ and $\boldsymbol{\Phi}^R_{i,t}$ can be replaced by a common STCM $\boldsymbol{\Phi}_{i,t}$,



(a) no structure (b) global RTF structure (c) ipsilateral RTF structure

Fig. 2. Illustration of the proposed spatial structures imposed on the speech STCVs, assuming $M = 2$ microphones per hearing device ($L = 1$, $R = 3$). The parameters to be estimated are highlighted in blue once per structure in order to emphasize the parameter reuse achieved by the global RTF structure and the ipsilateral RTF structure.

i.e.,

$$\mathbf{\Phi}_{i,t}^{L} = \mathbf{\Phi}_{i,t}^{R} =: \mathbf{\Phi}_{i,t}. \tag{33}$$

This assumption is generally more valid at lower signal-to-noise-ratios (SNRs), where the noise STCM becomes more dominant relative to the uncorrelated speech STCMs. As will be demonstrated in Section VI, this assumption holds quite well in practice. It should be noted that when combining this assumption with the global RTF structure—where the left and right speech STCVs are parallel—the resulting binaural STWF filter vectors are parallel as well.

C. Bilateral Correlation

The binaural STWF using the proposed speech STCV and interference STCM structures in Sections III-A and III-B exploits spatio-temporal correlation between both hearing devices, requiring the microphone signals to be transmitted between the left and right hearing devices. In order to investigate the performance benefit achieved by binaural processing, we will also consider bilateral processing, where both hearing devices operate independently. This corresponds to assuming no correlation between the left and right hearing devices for both the speech and the interference components such that the left and right speech STCVs can be modeled using a non-zero subvector for ipsilateral correlations and a zero subvector for contralateral correlations, i.e.,

$$\boldsymbol{\gamma}_{t,\text{bil}}^{L} = \begin{bmatrix} \bar{\gamma}_{t,1}^{L} \\ \vdots \\ \bar{\gamma}_{t,M}^{L} \\ \mathbf{0}_{MN\times 1} \end{bmatrix}, \quad \boldsymbol{\gamma}_{t,\text{bil}}^{R} = \begin{bmatrix} \mathbf{0}_{MN\times 1} \\ \bar{\gamma}_{t,M+1}^{R} \\ \vdots \\ \bar{\gamma}_{t,2M}^{R} \end{bmatrix}, \tag{34}$$

and, similarly, the common interference STCM in (33) can be modeled using non-zero submatrices for ipsilateral correlations and zero submatrices for the contralateral correlations, i.e.,

$$\mathbf{\Phi}_{i,t,\text{bil}} = \begin{bmatrix} \tilde{\mathbf{\Phi}}_{i,t}^{LL} & \mathbf{0}_{MN\times MN} \\ \mathbf{0}_{MN\times MN} & \tilde{\mathbf{\Phi}}_{i,t}^{RR} \end{bmatrix}. \tag{35}$$

Since

$$\mathbf{\Phi}_{i,t,\text{bil}}^{-1}\boldsymbol{\gamma}_{t,\text{bil}}^{L} = \begin{bmatrix} \left(\tilde{\mathbf{\Phi}}_{i,t}^{LL}\right)^{-1}[\boldsymbol{\gamma}_{t}^{L}]_{1:MN} \\ \mathbf{0}_{MN\times 1} \end{bmatrix} \tag{36}$$

$$\mathbf{\Phi}_{i,t,\text{bil}}^{-1}\boldsymbol{\gamma}_{t,\text{bil}}^{R} = \begin{bmatrix} \mathbf{0}_{MN\times 1} \\ \left(\tilde{\mathbf{\Phi}}_{i,t}^{RR}\right)^{-1}[\boldsymbol{\gamma}_{t}^{R}]_{MN+1:2MN} \end{bmatrix}, \tag{37}$$

the binaural STWF in (16) reduces to a set of bilateral filters, i.e.,

$$\mathbf{w}_{t,\text{bilat}}^{L} = \begin{bmatrix} [\mathbf{w}_{t}^{L}]_{1:MN} \\ \mathbf{0}_{MN\times 1} \end{bmatrix}, \quad \mathbf{w}_{t,\text{bilat}}^{R} = \begin{bmatrix} \mathbf{0}_{MN\times 1} \\ [\mathbf{w}_{t}^{R}]_{MN+1:2MN} \end{bmatrix}, \tag{38}$$

where $\mathbf{w}_{t,\text{bilat}}^{L}$ only depends on quantities related to the left hearing device and $\mathbf{w}_{t,\text{bilat}}^{R}$ only depends on quantities related to the right hearing device.

For the bilateral STWF, it is also possible to consider the ipsilateral RTF structure for the speech STCV in (34), i.e.,

$$\boldsymbol{\gamma}_{t,\text{bil,ipsi}}^{L} = \begin{bmatrix} \mathbf{h}_{t,\text{ipsi}}^{L} \otimes \bar{\gamma}_{t,L}^{L} \\ \mathbf{0}_{MN\times 1} \end{bmatrix}, \quad \boldsymbol{\gamma}_{t,\text{bil,ipsi}}^{R} = \begin{bmatrix} \mathbf{0}_{MN\times 1} \\ \mathbf{h}_{t,\text{ipsi}}^{R} \otimes \bar{\gamma}_{t,L}^{R} \end{bmatrix}, \tag{39}$$

with $\mathbf{h}_{t,\text{ipsi}}^{L}$ and $\mathbf{h}_{t,\text{ipsi}}^{R}$ defined in (28) and (29). As will be demonstrated in Section VI, these bilateral structures introduce relatively large estimation errors in practice, suggesting that the correlation between the left and right hearing devices for both the speech and the interference components is quite relevant.

IV. Deep Binaural Spatio-Temporal Wiener Filter

To estimate all required quantities, the binaural STWF is embedded into a supervised learning framework (see Fig. 3). Specifically, the speech STCVs $\boldsymbol{\gamma}_{t}^{L}$ and $\boldsymbol{\gamma}_{t}^{R}$, the speech PSDs ϕ_{t}^{L} and ϕ_{t}^{R}, as well as the inverse STCMs $(\mathbf{\Phi}_{i,t}^{L})^{-1}$ and $(\mathbf{\Phi}_{i,t}^{R})^{-1}$ are estimated using TCNs. Separate TCNs are used for the quantities related to the speech component (i.e., STCVs and PSDs) and for the quantities related to the interference component (i.e., inverse STCMs), where both TCNs are jointly trained using a loss function that compares the ground-truth binaural speech components to the estimated components obtained at the output of the deep binaural STWF. As input features, we utilized a concatenation of the logarithmic magnitude, as well as the cosine and sine of the phase, of the noisy STFT coefficients at all

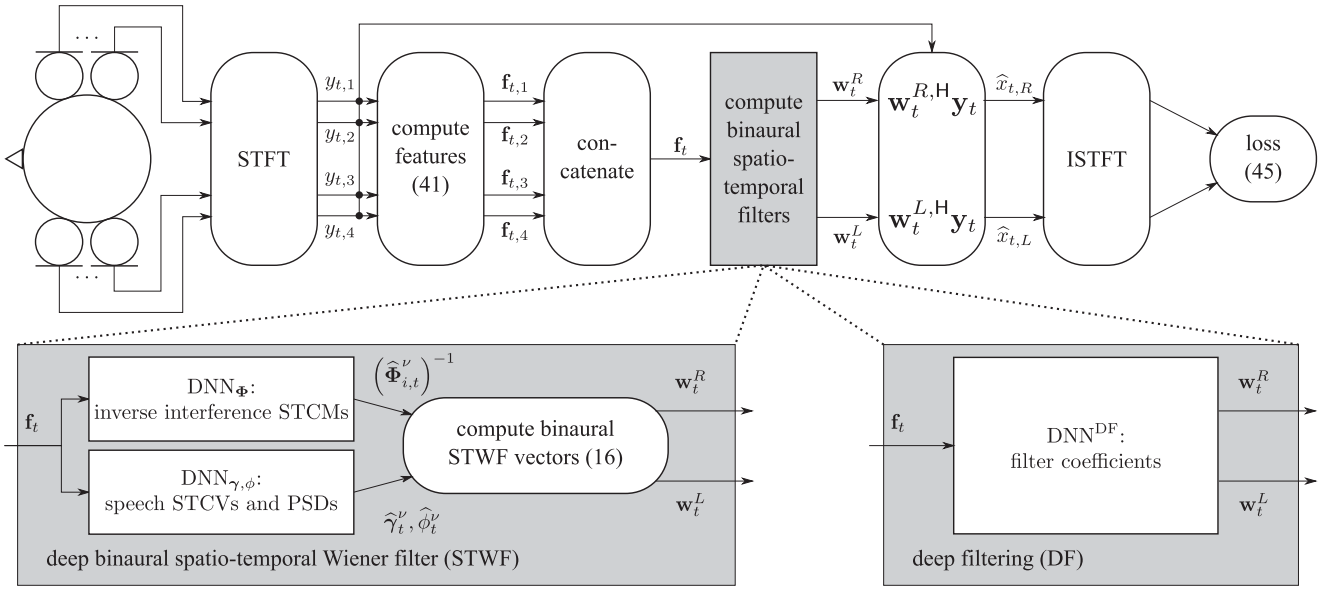


Fig. 3. Block diagram of the proposed deep binaural STWF and the baseline deep filtering algorithm, assuming $M = 2$ microphones per hearing device.

TABLE I
THE NUMBER OF REQUIRED PARAMETERS PER FREQUENCY BIN TO ESTIMATE THE SPEECH STCVS AND THE INVERSE INTERFERENCE STCMS FOR THE PROPOSED SPATIO-TEMPORAL CORRELATION STRUCTURES (ASSUMING $N = 5$ AND $M = 2$; $\nu \in \{l, R\}$), AS WELL AS THE MODEL MISMATCH ON THE EVALUATION DATASET IN TERMS OF THE MEAN RELATIVE ℓ_2 NORM ϵ_{ℓ_2}, THE MEAN HERMITIAN ANGLE ϵ_θ, THE MEAN RELATIVE FROBENIUS NORM ϵ_{FRO}, AND THE MEAN CORRELATION MATRIX DISTANCE ϵ_{CMD} IN (56)

quantity	structure	required parameters	ϵ_{ℓ_2}/dB (↓)	$\epsilon_\theta/^\circ$ (↓)	ϵ_{Fro}/dB (↓)	ϵ_{CMD} (↓)
$\boldsymbol{\gamma}_t^\nu$	—	$4(2MN-1) \triangleq$ 76	$-\infty$	0.0	—	—
	global RTF	$2(2M+N-2) \triangleq$ **14**	−4.5	30.1	—	—
	ipsilateral RTF	$4(M+2N-2) \triangleq$ 40	**−15.6**	**9.9**	—	—
	bilateral	$4(MN-1) \triangleq$ 36	−3.2	43.7	—	—
	bilateral & ipsilateral RTF	$4(M+N-2) \triangleq$ 20	−3.0	44.9	—	—
$\mathbf{\Phi}_{i,t}^\nu$	—	$8(MN)^2 \triangleq$ 800	—	—	$-\infty$	0.000
	common STCM ($\mathbf{\Phi}_{i,t}^L = \mathbf{\Phi}_{i,t}^R$)	$4(MN)^2 \triangleq$ 400	—	—	**−20.2**	**0.004**
	bilateral	$2(MN)^2 \triangleq$ **200**	—	—	−4.0	0.220

microphones[2], i.e.,

$$\mathbf{f}_{t,m} = \left[\log_{10}(|y_{t,m}|) \quad \cos(\angle y_{t,m}) \quad \sin(\angle y_{t,m})\right]^{\mathsf{T}} \tag{40}$$

$$\mathbf{f}_t = \left[\mathbf{f}_{t,1}^{\mathsf{T}} \quad \dots \quad \mathbf{f}_{t,2M}^{\mathsf{T}}\right]^{\mathsf{T}}, \tag{41}$$

where $\angle\cdot$ denotes the phase. We chose both the cosine and the sine of the phase to obtain an unambiguous and smooth phase representation [32]. For more details about the TCNs and the loss function, we refer to Section V-C. For the spatio-temporal correlation structures proposed in Section III, in the following we explain in more detail the parameters that are estimated by the TCNs. Table I provides an overview of the number of parameters per frequency bin for the speech STCVs and the inverse interference STCMs, which greatly differ between the different structures.

[2]In preliminary experiments, this feature choice outperformed the use of the real and imaginary parts of the noisy STFT coefficients as input features.

A. Speech Correlation Vectors and Power Spectral Densities

a) No Structure: When not imposing any structure on the speech STCVs, estimates of both complex-valued vectors $\boldsymbol{\gamma}_t^L$ and $\boldsymbol{\gamma}_t^R$ and both PSDs ϕ_t^L and ϕ_t^R are required. Since one element of each speech STCV is equal to 1 (cf. (10)) each speech STCV is determined by $2(2MN-1)$ real-valued parameters. Similarly to [33], the speech PSDs are estimated by applying real-valued masks $\widehat{a}_{t,L}$ and $\widehat{a}_{t,R}$ to the noisy STFT coefficients at the reference microphones, i.e.,

$$\widehat{\phi}_t^\nu = |\widehat{a}_{t,\nu} y_{t,\nu}|^2, \quad \widehat{a}_{t,\nu} \in [0,1], \quad \nu \in \{L, R\}, \tag{42}$$

with each mask determined by a single real-valued parameter, ensuring the range $[0, 1]$ with a sigmoid activation function. A single TCN uses the features in (40) at all microphones to estimate the undetermined parameters of the speech STCVs and the PSD masks.

b) Global RTF Structure: When imposing the global RTF structure on the speech STCVs, estimates of the global RTF vector $\mathbf{h}_{t,\text{glob}}$ in (22) and the speech temporal correlation vector $\bar{\boldsymbol{\gamma}}^{L}_{t,1}$ in (24) are required, separating the estimation process into a spatial factor and a temporal factor. The global RTF vector is determined by $2(2M-1)$ real-valued parameters, while the speech temporal correlation vector is determined by $2(N-1)$ real-valued parameters. Since the left and right speech PSDs are directly related by the squared magnitude of the global RTF (cf. (25)), only one PSD mask needs to be estimated for the global reference microphone. A single TCN uses the features in (40) at all microphones to estimate the undetermined parameters of the global RTF vector, the speech temporal correlation vector, and the PSD mask.

c) Ipsilateral RTF Structure: When imposing the ipsilateral RTF structure on the speech STCVs, estimates of the ipsilateral RTF vectors $\mathbf{h}^{\text{ipsi}}_{t,L}$ and $\mathbf{h}^{\text{ipsi}}_{t,R}$ in (28) and (29) as well as the speech temporal correlation vectors $\bar{\boldsymbol{\gamma}}^{L}_{t,L}$, $\bar{\boldsymbol{\gamma}}^{R}_{t,R}$, $\bar{\boldsymbol{\gamma}}^{R}_{t,L}$ and $\bar{\boldsymbol{\gamma}}^{L}_{t,R}$ in (31) are required. The ipsilateral RTF vectors are determined by $2(M-1)$ real-valued parameters each, the vectors $\bar{\boldsymbol{\gamma}}^{L}_{t,L}$ and $\bar{\boldsymbol{\gamma}}^{R}_{t,R}$ are determined by $2(N-1)$ real-valued parameters each, and the vectors $\bar{\boldsymbol{\gamma}}^{R}_{t,L}$ and $\bar{\boldsymbol{\gamma}}^{L}_{t,R}$ are determined by $2N$ real-valued parameters each (since none of the elements needs to be equal to 1). The ipsilateral RTF structure does not impose an explicit relationship between the left and right PSDs. A single TCN uses the features in (40) at all microphones to estimate the undetermined parameters of the ipsilateral RTF vectors, the speech temporal correlation vectors, and the PSD masks.

B. Interference Covariance Matrices

a) No Structure: We assume that the interference STCMs are full-rank, such that they are positive-definite, i.e., all eigenvalues are real-valued and larger than zero. Hence, the inverse interference STCMs are also positive-definite and can be decomposed using the Cholesky decomposition [15], [23], [26], [34] as

$$\left(\boldsymbol{\Phi}^{\nu}_{i,t}\right)^{-1} = \mathbf{L}^{\nu}_{i,t}\mathbf{L}^{\nu,\mathsf{H}}_{i,t}, \quad \nu \in \{L,R\}, \tag{43}$$

where the Cholesky factor $\mathbf{L}^{\nu}_{i,t} \in \mathbb{C}^{2MN\times 2MN}$ is a lower-triangular matrix with real-valued and positive diagonal elements, determined by $(2MN)^2$ real-valued parameters. A single TCN uses the features in (40) at all microphones to estimate these parameters. Similarly as in [15], [23], we then construct the Cholesky factors $\widehat{\mathbf{L}}^{\nu}_{i,t}$ by using disjoint subsets of the parameters for the real strictly lower triangular part, the imaginary strictly lower triangular part, and the real positive diagonal part, ensuring positivity of the diagonal part with a softplus activation function. Finally, we construct the inverse interference STCMs as in (43), i.e., without explicitly computing a matrix inverse.

b) Common Interference Covariance Matrix: When assuming the left and right interference STCMs to be equal, a single Cholesky factor and inverse interference STCM is estimated using the procedure described above, reducing the number of parameters by half to $(2MN)^2$.

C. Bilateral Correlation

When imposing a bilateral structure on the speech STCVs and the interference STCMs, only the ipsilateral correlations need to be estimated. The ipsilateral speech STCVs in (34) are determined by $2(MN-1)$ real-valued parameters each. When additionally imposing the ipsilateral RTF structure, the speech STCVs in (39) are determined by $2(M-1)$ real-valued parameters for each of the ipsilateral RTF vectors $\mathbf{h}^{L}_{t,\text{ipsi}}$ and $\mathbf{h}^{R}_{t,\text{ipsi}}$ as well as $2(N-1)$ real-valued parameters for each of the speech temporal correlation vectors $\bar{\boldsymbol{\gamma}}^{L}_{t,L}$ and $\bar{\boldsymbol{\gamma}}^{R}_{t,R}$. The Cholesky factor of the inverse submatrices $(\tilde{\boldsymbol{\Phi}}^{LL}_{i,t})^{-1}$ and $(\tilde{\boldsymbol{\Phi}}^{RR}_{i,t})^{-1}$ in (36) and (37) is determined by $(MN)^2$ real-valued parameters each. In contrast to the estimation procedures for the binaural STWF, separate TCNs are used for the left and right filters of the bilateral STWF. More specifically, one TCN uses the features in (40) at the microphones of the left hearing device to estimate the undetermined parameters of the left ipsilateral speech STCV and the left PSD mask, while another TCN uses the features at the microphones of the right hearing device to estimate the undetermined parameters of the right ipsilateral speech STCV and the right PSD mask. Similarly, two separate TCNs are used to estimate the undetermined parameters of the left and right inverse interference STCMs.

V. Simulation Setup

In this section, we present our simulation setup, consisting of the used datasets (Section V-A), the baseline binaural speech enhancement algorithms (Section V-B), and the settings of all algorithms (Section V-C).

A. Datasets

To train, validate, and evaluate all supervised learning-based binaural speech enhancement algorithms, we constructed datasets using diverse speech and noise source material from the DNS 1 and DNS 2 challenge datasets [27], [28] and simulated BRIRs from the CEC 1 dataset [29]. In addition, to test the generalization capabilities of the considered algorithms in more realistic scenarios, we considered the CEC 3 dataset (Task 3)[3]. All datasets were used at a sampling rate of 16 kHz.

1) Training and Validation: For the training and validation datasets, we used the speech and noise source material from the DNS 2 challenge dataset, consisting of English sentences from 11350 speakers and 600 noise classes. We chose not to use the speech and noise source material from the CEC 1 dataset in order to increase speaker and noise diversity. For the BRIRs, we used the CEC 1 training dataset, consisting of 6000 different room configurations. These BRIRs were simulated for a randomly positioned directional speech source and a randomly positioned omnidirectional noise source captured by binaural behind-the-ear hearing aids mounted on an artificial head in randomly sized rooms with reverberation time T_{60} ranging from 0.2 s to 0.4 s (i.e., low to moderate reverberation). The hearing aids consisted

[3]https://claritychallenge.org/docs/cec3/task_3/cec3_task3_overview

of three microphones each (front, mid, and rear), with a microphone spacing of about 7.6 mm. The front and mid microphones were chosen for all simulations, i.e., $M = 2$ microphones were used per hearing device. The speech source was located at an angle within $+-30^\circ$ w.r.t. the listener, while the noise source could be positioned anywhere in the room except for less than 1 m from the walls or the listener. Surface absorption coefficients were varied to simulate various room characteristics such as doors, windows, curtains, rugs, or furniture. Random speech and noise sources were convolved with BRIRs corresponding to a randomly chosen room configuration before being mixed at better-ear SNRs ranging from 0 dB to 15 dB (considering the reference microphones at both hearing aids). In total, the training and validation datasets have a length of 80 h and 20 h, respectively, with each utterance of length 4 s.

2) Evaluation: For evaluation, we considered a matched dataset resembling the training dataset and a mismatched dataset designed to test generalization capabilities. For the matched evaluation dataset, we used speech and noise source material from the DNS 1 challenge evaluation dataset and BRIRs from the CEC 1 validation dataset. Random speech and noise sources were convolved with BRIRs corresponding to a randomly chosen room configuration before being mixed at better-ear SNRs from -5 dB to 20 dB in steps of 5 dB. The training/validation and matched evaluation datasets were disjoint in terms of speakers, noise sources, and BRIRs. In total, 100 utterances were considered per SNR, with each utterance of length 10 s.

For the mismatched evaluation dataset, we used a subset of the CEC 3 development dataset (Task 3), comprising real noise backgrounds and higher-order ambisonic RIRs recorded in complex environments (busy roads, railway platforms, and social gatherings) as well as simulated head rotation. Note that we omitted the social gatherings environment, which would have required a target speech extraction approach and is thus incompatible with the approach in this paper. Random speech sources were convolved with higher-order ambisonic RIRs, rotated to simulate head movement, and binauralized with measured head-related transfer functions. The resulting speech components were mixed with real noise backgrounds at better-ear SNRs from -5 dB to 6 dB. This SNR range represents a subset of the full -12 dB to 6 dB range, chosen to provide a reasonable mismatch with our training dataset (which contained the range from 0 dB to 15 dB). The considered evaluation dataset presents mismatches in terms of speakers, noise types, noise spatial coherence, acoustic conditions (as reflected in the recorded RIRs), and the dynamic aspect of head rotation. In total, 997 utterances were selected, with each utterance lasting about 5 s.

B. Baseline Algorithms

We consider three baseline binaural speech enhancement algorithms, where two algorithms are causal and one algorithm is non-causal. The first baseline algorithm is causal and directly estimates the binaural multi-frame filter coefficients in the STFT domain (see Fig. 3), which can be viewed as a binaural extension of the deep filtering (DF) algorithm proposed in [16]. More specifically, rather than estimating the speech STCVs, speech PSDs, and inverse interference STCMs to compute the binaural STWF vectors using (16), the DF algorithm directly estimates the binaural spatio-temporal filter vectors $\mathbf{w}_t^L$ and $\mathbf{w}_t^R \in \mathbb{C}^{2MN}$, each determined by $4MN$ real-valued parameters. A single TCN uses the features in (40) at all microphones to estimate these parameters, ensuring the range $[-1, 1]$ with a hyperbolic tangent activation function as proposed in [16].

As second state-of-the-art baseline algorithm, we consider the causal binaural Conv-TasNet algorithm, which uses a learned transform instead of the STFT and a TCN-based separator that estimates real-valued masks employed in a mask-and-sum approach [9].

As third state-of-the-art baseline algorithm, we use the non-causal BCCTN algorithm [14], which uses a complex-valued convolutional transformer network that estimates complex-valued time-frequency masks for the left and right reference channels. Note that the BCCTN algorithm uses a non-causal multi-head attention implementation and is thus not suitable for real-time processing—in contrast to all other considered algorithms.

C. Algorithmic Settings

We used the same STFT framework and input features for the STWF algorithms and the DF algorithm. To increase speech correlation across successive STFT frames, we used a high temporal resolution, i.e., a frame length of 8 ms and 75 % overlap, resulting in a low input-output latency. We used a square-root Hann window as analysis and synthesis window. As multi-frame filter length, the STWF algorithms and the DF algorithm used $N = 5$ frames, such that temporal correlations within 16 ms could be exploited. This choice represents a trade-off between capturing sufficient temporal context to exploit speech correlations and maintaining reasonable computational complexity. To limit speech distortion, a minimum gain of $g_{\min} = -20$ dB was applied to the binaural output signals of the STWF algorithms and the DF algorithm during evaluation. The final estimated binaural target speech components were thus obtained as

$$\widehat{x}_{t,\nu} = \begin{cases} g_{\min} y_{t,\nu}, & \text{if } |\mathbf{w}_t^{\nu,\mathsf{H}} \mathbf{y}_t| < |g_{\min} y_{t,\nu}| \\ \mathbf{w}_t^{\nu,\mathsf{H}} \mathbf{y}_t, & \text{else} \end{cases}. \tag{44}$$

To train all algorithms except the BCCTN algorithm, we used the following STFT-domain loss function proposed in [35]:

$$L_{f,t,\nu}^b = \beta \left| x_{f,t,\nu}^b - \widehat{x}_{f,t,\nu}^b \right| + (1-\beta) \left| |x_{f,t,\nu}^b| - |\widehat{x}_{f,t,\nu}^b| \right|$$

$$L = \frac{1}{2BFT} \sum_{b=1}^{B} \sum_{f=1}^{F} \sum_{t=1}^{T} \sum_{\nu \in \{L,R\}} L_{f,t,\nu}^b, \tag{45}$$

where $B = 4$ denotes the batch size, b denotes the batch element, F and T denote the numbers of frequency bins and frames in an utterance, and $\beta = 0.4$ is a hyperparameter chosen as in [35]. The magnitude term helps preserve spectral shape and formant structure, while the complex-valued difference term helps preserve phase relationships. Note that this loss was computed after the estimated binaural target speech components in (5) were transformed back to the time domain using an inverse STFT (see Fig. 3), followed by an additional transformation to the STFT

domain with a frame length of $32\,\mathrm{ms}$ and $50\,\%$ overlap. For the BCCTN algorithm, we used the STFT framework and loss function proposed in [14], which incorporates terms reflecting noise reduction, intelligibility improvement, and interaural cue preservation.

We used TCNs as the DNN architecture for the STWF algorithms and the DF algorithm, implemented based on the official (monaural) Conv-TasNet implementation.[4] We fixed the number of stacks to two, the number of layers to six, the kernel size to three, and the bottleneck size to 32, yielding a temporal receptive field size of $512\,\mathrm{ms}$. All STWF algorithms (imposing different correlation structures) and the DF algorithm were trained separately, with the TCN architecture remaining identical and only adapting its final layer to match the required parameter count for each algorithm (presented in Table I for the STWF algorithms).

For both the binaural Conv-TasNet algorithm and the BCCTN algorithm, we used the code provided by the authors and the DNN hyperparameters proposed in [9] and [14], respectively.

All algorithms were trained for a maximum of 100 epochs with early stopping using the AdamW optimizer [36] and with gradient ℓ_2 norms clipped to 5. The learning rate was initialized at 10^{-3} and halved after three epochs without validation loss improvement. Early stopping was applied after ten epochs without validation loss improvement. For the STWF algorithms and the DF algorithm, complex-valued numbers were constructed from the (real-valued) TCN outputs by assigning separate output elements for the real and imaginary parts. The simulations were implemented using PyTorch 2.0.1 [37] and executed on NVIDIA GeForce RTX A5000 graphics cards.

VI. Validity of Spatio-Temporal Correlation Structures

In this section, we describe an approach to validate the proposed spatio-temporal correlation structures for the speech STCVs and interference STCMs in Section III. We first discuss how to compute ground-truth STCVs and STCMs, impose spatio-temporal structure on these quantities, introduce several metrics to evaluate the incurred model mismatch, and finally present the validation results.

A. Ground-Truth STCVs and STCMs

To compute the ground-truth speech STCVs and interference STCMs, we first apply recursive smoothing on the instantaneous oracle speech and noise STCMs, i.e.,

$$\mathbf{\Phi}_{x,t} = \alpha\mathbf{\Phi}_{x,t-1} + (1-\alpha)\mathbf{x}_t\mathbf{x}_t^{\mathsf{H}} \tag{46}$$

$$\mathbf{\Phi}_{n,t} = \alpha\mathbf{\Phi}_{n,t-1} + (1-\alpha)\mathbf{n}_t\mathbf{n}_t^{\mathsf{H}}. \tag{47}$$

To track rapidly varying speech and noise statistics, we set the smoothing constant α to match the frame shift of the employed STFT (equal to $2\,\mathrm{ms}$), using the relation $\alpha = \exp(-T_s/\tau)$, where $T_s = 2\,\mathrm{ms}$ denotes the frame shift and $\tau = 2\,\mathrm{ms}$ denotes the smoothing time constant. Using (46), the left and right speech STCVs and PSDs are computed using (9), i.e.,

$$\boldsymbol{\gamma}_{t,\nu} = \frac{\mathbf{\Phi}_{x,t}\mathbf{e}_\nu}{\phi_t^\nu}, \qquad \nu \in \{L,R\} \tag{48}$$

$$\phi_t^\nu = \mathbf{e}_\nu^{\mathsf{T}}\mathbf{\Phi}_{x,t}\mathbf{e}_\nu, \qquad \nu \in \{L,R\}. \tag{49}$$

The left and right interference STCMs are computed using (13), i.e.,

$$\mathbf{\Phi}_{i,t}^\nu = \mathbf{\Phi}_{x,t} - \phi_t^\nu\boldsymbol{\gamma}_t^\nu\boldsymbol{\gamma}_t^{\nu,\mathsf{H}} + \mathbf{\Phi}_{n,t}, \quad \nu \in \{L,R\}. \tag{50}$$

B. Spatio-Temporal Structures

To impose the global RTF structure or the ipsilateral RTF structure (see Section III-A) on the ground-truth speech STCVs, the ground-truth RTFs are first computed according to (18) using the oracle speech STCM (46) and the oracle speech PSDs (49). For the global RTF structure, the speech STCVs are constructed using (23) and (26), where the global RTF vector $\mathbf{h}_{t,\mathrm{glob}}$ in (22) is constructed using the ground-truth RTFs and the speech temporal correlation vector $\bar{\boldsymbol{\gamma}}_{t,1}^L$ is extracted from the speech STCV in (48). Similarly, for the ipsilateral RTF structure, the speech STCVs are constructed using (30), where the ipsilateral RTF vectors $\mathbf{h}_{t,\mathrm{ipsi}}^L$ and $\mathbf{h}_{t,\mathrm{ipsi}}^R$ in (28) and (29) are constructed using the ground-truth RTFs, and the speech temporal correlation vectors $\bar{\boldsymbol{\gamma}}_{t,L}^\nu$ and $\bar{\boldsymbol{\gamma}}_{t,R}^\nu$ are extracted from the speech STCV in (48).

The common interference STCM (see Section III-B) is constructed as the matrix minimizing the squared Frobenius norm of the difference with the ground-truth interference STCMs $\mathbf{\Phi}_{i,t}^L$ and $\mathbf{\Phi}_{i,t}^R$ in (50), i.e.,

$$\tilde{\mathbf{\Phi}}_{i,t} = \mathrm{argmin}_{\mathbf{\Phi}} \left(\left\| \mathbf{\Phi} - \mathbf{\Phi}_{i,t}^L \right\|_F^2 + \left\| \mathbf{\Phi} - \mathbf{\Phi}_{i,t}^R \right\|_F^2 \right) \tag{51}$$

$$= \frac{1}{2}\left(\mathbf{\Phi}_{i,t}^L + \mathbf{\Phi}_{i,t}^R \right). \tag{52}$$

For the bilateral correlation structures (see Section III-C), the coefficients of the speech STCVs and the interference STCMs corresponding to the contralateral correlations are simply set to zero.

C. Results

We evaluate the model mismatch incurred by imposing spatio-temporal correlation structures on the ground-truth speech STCVs and interference STCMs in terms of several metrics on the evaluation dataset. To evaluate model mismatch for the speech STCVs, we consider the mean relative ℓ_2 norm and the mean Hermitian angle, defined as

$$\epsilon_{\ell_2} = \frac{1}{2FT}\sum_{f=1}^{F}\sum_{t=1}^{T}\sum_{\nu\in\{L,R\}} \frac{\left\| \tilde{\boldsymbol{\gamma}}_{f,t}^\nu - \boldsymbol{\gamma}_{f,t}^\nu \right\|_2}{\left\| \boldsymbol{\gamma}_{f,t}^\nu \right\|_2} \tag{53}$$

$$\epsilon_\theta = \frac{1}{2FT}\sum_{f=1}^{F}\sum_{t=1}^{T}\sum_{\nu\in\{L,R\}} \arccos\left(\frac{|\tilde{\boldsymbol{\gamma}}_{f,t}^{\nu,\mathsf{H}}\boldsymbol{\gamma}_{f,t}^\nu|}{\left\| \tilde{\boldsymbol{\gamma}}_{f,t}^\nu \right\|_2 \left\| \boldsymbol{\gamma}_{f,t}^\nu \right\|_2} \right), \tag{54}$$

[4] https://github.com/naplab/Conv-TasNet

where $\boldsymbol{\gamma}_{f,t}^{\nu}$ corresponds to the ground-truth speech STCVs in (48) and $\tilde{\boldsymbol{\gamma}}_{f,t}^{\nu}$ corresponds to the speech STCVs with imposed structure.

To evaluate model mismatch for the interference STCMs, we consider the mean relative Frobenius norm and the mean correlation matrix distance [38], defined as

$$\epsilon_{\text{Fro}} = \frac{1}{2FT}\sum_{f=1}^{F}\sum_{t=1}^{T}\sum_{\nu\in\{L,R\}}\frac{\left\|\tilde{\boldsymbol{\Phi}}_{i,f,t}-\boldsymbol{\Phi}_{i,f,t}^{\nu}\right\|_F}{\left\|\boldsymbol{\Phi}_{i,f,t}^{\nu}\right\|_F} \tag{55}$$

$$\epsilon_{\text{CMD}} = \frac{1}{2FT}\sum_{f=1}^{F}\sum_{t=1}^{T}\sum_{\nu\in\{L,R\}}\left(1-\frac{\text{trace}\{\tilde{\boldsymbol{\Phi}}_{i,f,t}\boldsymbol{\Phi}_{i,f,t}^{\nu,\mathsf{H}}\}}{\|\tilde{\boldsymbol{\Phi}}_{i,f,t}\|_F\|\boldsymbol{\Phi}_{i,f,t}^{\nu}\|_F}\right), \tag{56}$$

where $\boldsymbol{\Phi}_{i,f,t}^{\nu}$ corresponds to the ground-truth interference STCMs in (50) and $\tilde{\boldsymbol{\Phi}}_{i,f,t}$ corresponds to the common interference STCM in (51). The correlation matrix distance ϵ_{CMD} can be interpreted as the angle between the vectorized STCMs $\tilde{\boldsymbol{\Phi}}_{i,f,t}$ and $\boldsymbol{\Phi}_{i,f,t}$ in $(2MN)^2$-dimensional space, yielding values between 0 and 1, where smaller values denote higher similarity.

For all considered spatio-temporal correlation structures, Table I shows the number of required parameters to estimate the speech STCVs and the interference STCMs, both as a function of the number of microphones M and the multi-frame filter length N as well as specifically for $M = 2$ and $N = 5$ (used in the simulations). It also shows the model mismatch by imposing spatio-temporal correlation structures on the ground-truth speech STCVs and interference STCMs in terms of the metrics introduced in the previous subsection.

For the speech STCVs, it can be observed on the one hand that all spatio-temporal correlation structures significantly reduce the number of required parameters. The global RTF structure yields the largest reduction (factor 5.4), while the ipsilateral RTF structure yields the smallest reduction (factor 1.9). On the other hand, both the global RTF structure and the bilateral structures incur a relatively large model mismatch in terms of both metrics. Although the global RTF structure is commonly used in single-frame algorithms (typically utilizing longer STFT frames), the global RTF structure may be less suitable for the considered multi-frame algorithms, which rely on short STFT frames to exploit the temporal correlations of speech signals. In addition, due to the relatively large distance between the microphones on the left and right hearing devices, the speech STFT coefficients of the global reference microphone (on the left hearing device) may not be fully correlated with the speech STFT coefficients of the contralateral microphones (on the right hearing device). In contrast, the ipsilateral RTF structure incurs a smaller model mismatch while reducing the number of required parameters with a similar factor as the bilateral RTF structure.

For the interference STCMs, it can be observed on the one hand that assuming a common STCM reduces the number of required parameters by a factor 2, while the bilateral structure further reduces the number of required parameters by an additional factor 2. In absolute numbers, the parameter reductions for the interference STCMs are much larger than the reductions for the speech STCVs. On the other hand, the bilateral interference correlation structure incurs a relatively large model mismatch in terms of both metrics, consistent with the model mismatch for the bilateral speech correlation structure. In contrast, assuming a common interference STCM incurs a smaller model mismatch. The observations in terms of model mismatch for the considered spatio-temporal correlation structures will be confirmed by the simulation results in the next section.

VII. Simulation Results

In this section, we investigate the speech enhancement performance and the computational complexity of the deep binaural STWF and the deep bilateral STWF using the proposed correlation structures. Furthermore, we compare their performance with three baseline algorithms, i.e., the binaural DF algorithm, the binaural Conv-TasNet algorithm, and the BCCTN algorithm. As mentioned before, all considered algorithms except the BCCTN algorithm are causal. To evaluate computational complexity, we consider the RF, defined as processing duration vs. utterance duration using a single thread on an AMD EPYC 7443P CPU, as well as the number of MACS (determined using the PyTorch profiler) and the number of trainable weights. To evaluate speech enhancement performance, we consider the wideband perceptual evaluation of speech quality (PESQ) [39] metric and the hearing aid speech quality index (HASQI) [40] as objective metrics of speech quality, as well as the hearing aid speech perception index (HASPI) [41] as an objective metric of speech intelligibility. To evaluate interaural cue preservation, we consider the interaural level difference (ILD) and interaural phase difference (IPD) errors between the output signals and the target speech components of the input signals. We use the ILD and IPD errors defined in [14], where both errors are computed only in STFT bins with active speech. The ILD error is defined as

$$\Delta\text{ILD} = \frac{1}{A}\sum_{f=1}^{F}\sum_{t=1}^{T}\mathcal{M}_{f,t}|\text{ILD}_{f,t}^{\text{out}}-\text{ILD}_{f,t}^{\text{in}}|, \tag{57}$$

where $\mathcal{M}_{f,t}$ denotes an ideal binary mask indicating STFT bins with active speech computed from the target speech components (see [14] for more details), and $A = \sum_f^F\sum_t^T\mathcal{M}_{f,t}$ denotes the number of STFT bins with active speech. The ILDs between the output signals and the ILDs between the target speech components of the input signals are defined as

$$\text{ILD}_{f,t}^{\text{out}} = \frac{|\widehat{x}_{f,t,L}|^2}{|\widehat{x}_{f,t,R}|^2}, \quad \text{ILD}_{f,t}^{\text{in}} = \frac{|x_{f,t,L}|^2}{|x_{f,t,R}|^2}. \tag{58}$$

Similarly, the IPD error is defined as

$$\Delta\text{IPD} = \frac{1}{A}\sum_{f=1}^{F}\sum_{t=1}^{T}\mathcal{M}_{f,t}|\text{IPD}_{f,t}^{\text{out}}-\text{IPD}_{f,t}^{\text{in}}|, \tag{59}$$

where the IPDs between the output signals and the IPDs between the target speech components of the input signals are defined as

$$\text{IPD}_{f,t}^{\text{out}} = \angle\left(\frac{\widehat{x}_{f,t,L}}{\widehat{x}_{f,t,R}}\right), \quad \text{IPD}_{f,t}^{\text{in}} = \angle\left(\frac{x_{f,t,L}}{x_{f,t,R}}\right). \tag{60}$$

TABLE II
COMPUTATIONAL COMPLEXITY IN TERMS OF THE AVERAGE REAL-TIME FACTOR (RF), THE NUMBER OF MULTIPLY-ACCUMULATE OPERATIONS PER SECOND (MACS), AND THE NUMBER OF TRAINABLE WEIGHTS FOR THE DEEP BINAURAL STWF AND THE DEEP BILATERAL STWF (IMPOSING DIFFERENT CORRELATION STRUCTURES) AS WELL AS THE BINAURAL DF ALGORITHM, THE BINAURAL CONV-TASNET ALGORITHM, AND THE NON-CAUSAL BCCTN ALGORITHM

	$\mathbf{\Phi}_{i,t}^{L} = \mathbf{\Phi}_{i,t}^{R}$	γ_t^{ν} structure	RF / % (↓)	MACS / M (↓)	trainable weights / M (↓)
noisy	—	—	—	—	—
binaural STWF	✗	✗	54.6	539	2.10
	✓	✗	36.2	287	1.30
	✓	global RTF	25.1	273	1.19
	✓	ipsilateral RTF	35.0	287	1.24
bilateral STWF	—	✗	12.1	76	0.42
	—	ipsilateral RTF	12.4	76	0.41
binaural DF	—	—	**4.8**	**6**	**0.34**
binaural Conv-TasNet [9]	—	—	36.0	75	1.67
BCCTN (non-causal) [14]	—	—	—	5730	11.09

TABLE III
SPEECH ENHANCEMENT PERFORMANCE IN TERMS OF AVERAGE PESQ, HASQI, AND HASPI VALUES AND INTERAURAL CUE PRESERVATION IN TERMS OF AVERAGE ILD AND IPD ERRORS FOR THE DEEP BINAURAL STWF AND THE DEEP BILATERAL STWF (IMPOSING DIFFERENT CORRELATION STRUCTURES) AS WELL AS THE BINAURAL DF ALGORITHM, THE BINAURAL CONV-TASNET ALGORITHM, AND THE NON-CAUSAL BCCTN ALGORITHM ON THE MATCHED EVALUATION DATASET

	$\mathbf{\Phi}_{i,t}^{L} = \mathbf{\Phi}_{i,t}^{R}$	γ_t^{ν} structure	PESQ (↑)	HASQI (↑)	HASPI (↑)	ΔILD / dB (↓)	ΔIPD / rad (↓)
noisy	—	—	1.62	0.39	0.90	—	—
binaural STWF	✗	✗	2.40	**0.50**	**0.95**	3.26	0.72
	✓	✗	2.38	**0.50**	**0.95**	3.27	0.73
	✓	global RTF	2.34	0.49	0.94	4.03	0.90
	✓	ipsilateral RTF	2.39	**0.50**	**0.95**	3.29	0.73
bilateral STWF	—	✗	2.19	0.47	0.94	4.14	0.78
	—	ipsilateral RTF	2.18	0.47	0.94	4.13	0.78
binaural DF	—	—	2.23	0.47	0.94	3.42	0.75
binaural Conv-TasNet [9]	—	—	2.19	0.49	0.94	3.86	0.81
BCCTN (non-causal) [14]	—	—	**2.48**	0.48	0.91	**2.62**	**0.66**

For all speech enhancement and interaural cue preservation metrics, we used the left and right reverberant speech signals at the reference microphones as the reference signals. For PESQ, we averaged the values across the left and right output signals, whereas HASQI and HASPI inherently consider only the better ear. For HASQI and HASPI, we assumed normal hearing (i.e., a flat hearing loss of 0 dB). Audio demos for matched and mismatched conditions (including a condition with a moving source entirely mismatched from the training dataset) can be found online.[5]

A. Computational Complexity

Table II shows the computational complexity for all considered binaural speech enhancement algorithms, where all metrics were averaged across all utterances and SNRs of the matched evaluation dataset. The DF algorithm achieves the lowest computational complexity, with an RF of 4.8 %, 6 M MACS, and 0.34 M trainable weights. In contrast, the non-causal BCCTN algorithm shows the highest computational complexity with 5730 M MACS and 11.09 M trainable weights, highlighting its substantial computational demand. The deep binaural STWF not imposing any correlation structure results in an RF of 54.6 %, 539 M MACS, and 2.10 M trainable weights. Imposing a common interference STCM reduces the RF to 36.2 %, while also decreasing the MACS to 287 M and the trainable weights to 1.30 M. Imposing the global RTF structure further reduces the RF to 25.1 %, the MACS to 273 M, and the trainable weights to 1.19 M. In contrast, imposing the ipsilateral RTF structure yields only minimal additional computational savings over the common interference STCM. Compared to the deep binaural STWF algorithms, the deep bilateral STWF algorithms result in lower computational complexity, with RFs around 12 %, 76 M MACS, and between 0.41 M and 0.42 M trainable weights.

B. Matched Evaluation Dataset

For the matched evaluation dataset, Table III shows the speech enhancement and interaural cue preservation performance for all considered binaural speech enhancement algorithms, averaged across all utterances and SNRs. First, it can be observed that all algorithms yield improvements in terms of all considered speech enhancement metrics. The deep binaural STWF not imposing any correlation structure achieves a high PESQ value (2.40) and the highest HASQI and HASPI values (0.50 and 0.95, respectively), while the non-causal BCCTN algorithm achieves the highest PESQ value (2.48), but only moderate HASQI and HASPI values (0.48, and 0.91). Compared to the unstructured variant, imposing a common interference STCM results in only minor PESQ reductions (2.38) while achieving the same HASQI and HASPI values. Further imposing the global RTF structure slightly reduces speech enhancement performance (PESQ of

[5]https://uol.de/en/sigproc/research/audio-demos/binaural-noise-reduction/stwf

TABLE IV
SPEECH ENHANCEMENT PERFORMANCE IN TERMS OF AVERAGE PESQ, HASQI, AND HASPI VALUES AND INTERAURAL CUE PRESERVATION IN TERMS OF AVERAGE ILD AND IPD ERRORS FOR THE DEEP BINAURAL STWF AND THE DEEP BILATERAL STWF (IMPOSING DIFFERENT CORRELATION STRUCTURES) AS WELL AS THE BINAURAL DF ALGORITHM, THE BINAURAL CONV-TASNET ALGORITHM, AND THE NON-CAUSAL BCCTN ALGORITHM ON THE MISMATCHED EVALUATION DATASET

	$\mathbf{\Phi}_{i,t}^{L} = \mathbf{\Phi}_{i,t}^{R}$	$\boldsymbol{\gamma}_{t}^{\nu}$ structure	PESQ (↑)	HASQI (↑)	HASPI (↑)	ΔILD / dB (↓)	ΔIPD / rad (↓)
noisy	—	—	1.09	0.07	0.42	—	—
binaural STWF	✗	✗	1.21	**0.12**	**0.55**	1.33	6.35
	✓	✗	1.20	0.11	0.54	1.33	6.35
	✓	global RTF	1.19	0.11	0.52	1.42	6.84
	✓	ipsilateral RTF	1.19	0.11	0.54	1.33	6.32
bilateral STWF	—	✗	1.18	0.11	0.54	1.40	7.93
	—	ipsilateral RTF	1.16	0.11	0.54	1.44	8.39
binaural DF	—	—	1.19	**0.12**	0.53	1.31	6.41
binaural Conv-TasNet [9]	—	—	1.14	0.09	0.44	1.44	7.45
BCCTN (non-causal) [14]	—	—	**1.23**	0.10	0.42	**1.07**	**4.69**

2.34). In contrast, imposing the ipsilateral RTF structure preserves the speech enhancement performance of the unstructured variant (PESQ of 2.39). The deep bilateral STWF algorithms, the binaural DF algorithm, and the binaural Conv-TasNet algorithm result in lower speech enhancement performance (PESQ around 2.20). The lower performance of the deep bilateral STWF algorithms is presumably caused by the lack of information exchange between the left and right hearing devices, which limits both the spatial diversity the TCNs can exploit as well as the number of microphones available for filtering.

In terms of interaural cue preservation, it can be observed that the non-causal BCCTN algorithm achieves the best performance, while the deep bilateral STWF algorithms and the deep binaural STWF algorithm imposing the global RTF structure exhibit the worst performance. Notably, imposing the ipsilateral RTF structure preserves interaural cues as effectively as the unstructured variant. The good interaural cue preservation of the BCCTN algorithm is presumably caused by the inclusion of loss terms that penalize interaural cue distortion. However, it should be noted that, from a perceptual perspective, all algorithms except the deep bilateral STWF algorithms preserve the binaural localization cues quite well.

C. Mismatched Evaluation Dataset

For the mismatched evaluation dataset, Table IV shows the speech enhancement and interaural cue preservation performance for all considered binaural speech enhancement algorithms, averaged across all utterances. Compared to the matched evaluation dataset, the speech enhancement performance of all algorithms is reduced, which is expected due to the strong mismatch between training and evaluation conditions. Nevertheless, all algorithms still yield improvements in terms of all considered speech enhancement metrics (except for the BCCTN algorithm in terms of HASPI). The speech enhancement and interaural cue preservation tendencies across algorithms are similar to those observed on the matched evaluation dataset, however with smaller differences between algorithms.

To experimentally evaluate the robustness of the considered binaural speech enhancement algorithms, we have also included an audio example of an acoustic scene that is entirely mismatched from the training dataset (featuring a moving target speaker in an unseen room, with unseen quasi-diffuse noise, and a hearing aid configuration with unseen inter-microphone spacings) on the webpage.[5]

D. Summary

Overall, the simulation results show that imposing spatio-temporal correlation structure on the speech STCVs and the interference STCMs reduces computational complexity in terms of the RF, MACS, and the number of trainable weights, while hardly affecting speech enhancement performance. Among the binaural STWF algorithms, the variant assuming a common interference STCM and the ipsilateral RTF structure offers the best trade-off in terms of computational complexity, speech enhancement performance, and interaural cue preservation.

VIII. CONCLUSIONS

In this paper, we proposed several procedures to impose spatio-temporal correlation structures on the speech STCVs and interference STCMs, required to implement the binaural STWF. These procedures mainly differ in terms of the relation between the microphones, particularly between the left and the right hearing device, as well as the number of parameters to be estimated. First, assuming that the spatial correlation of the speech component is stationary over the length of the multi-frame filter, we proposed to decompose the speech STCV as the Kronecker product of a spatial RTF vector and a temporal correlation vector. We either considered a single global reference microphone or a reference microphone for each hearing device. Second, we proposed to replace the left and right interference STCMs by a common interference STCM. In addition, we considered a bilateral STWF by neglecting all spatio-temporal correlations between both hearing devices. All required parameters were estimated by embedding the binaural STWF into a supervised learning framework.

Simulation results using both simulated and measured BRIRs as well as diverse speech and noise sources demonstrate that the combination of the speech STCV structure considering two reference microphones and a common interference STCM yields the best overall performance, reducing the real-time factor by around 36 % while maintaining speech enhancement and interaural cue preservation performance compared to not imposing

any spatio-temporal correlation structure. These results are consistent with a validation based on ground-truth quantities. Furthermore, the best deep binaural STWF algorithm outperforms two state-of-the-art binaural speech enhancement algorithms based on supervised learning, namely the deep filtering algorithm and the binaural Conv-TasNet algorithm, while approaching the performance of the (non-causal) BCCTN algorithm.

REFERENCES

[1] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: MIT Press, 1997.

[2] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," in *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 18–30, Mar., 2015.

[3] S. Gannot, E. Vincent, S. M.-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr., 2017. [Online]. Available: https://ieeexplore.ieee.org/document/7805139/

[4] S. S. Doclo Gannot, D. Marquardt, and E. Hadad, "Binaural speech processing with application to hearing devices," in *Proc. Audio Source Separation Speech*, 2018, pp. 413–442. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119279860.ch18

[5] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "A convex approximation of the relaxed binaural beamforming optimization problem," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 27, no. 2, pp. 321–331, Feb., 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8514022

[6] J. Zhang and G. Zhang, "A parametric unconstrained beamformer based binaural noise reduction for assistive hearing," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 30, pp. 292–304, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9664313

[7] S. Thaleiser and G. Enzner, "Binaural-projection multichannel Wiener filter for cue-preserving binaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 31, pp. 3730–3745, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10256053/

[8] X. Sun, R. Xia, J. Li, and Y. Yan, "A deep learning based binaural speech enhancement approach with spatial cues preservation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 5766–5770.

[9] C. Han, Y. Luo, and N. Mesgarani, "Real-time binaural speech separation with preserved spatial cues," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2020, pp. 6404–6408.

[10] Z. Sun, Y. Li, H. Jiang, F. Chen, X. Xie, and Z. Wang, "A supervised speech enhancement method for smartphone-based binaural hearing aids," in *IEEE Trans. Biomed. Circuits Syst.*, vol. 14, no. 5, pp. 951–960, Oct., 2020.

[11] B. J. Borgström, M. S. Brandstein, G. A. Ciccarelli, T. F. Quatieri, and C. J. Smalt, "Speaker separation in realistic noise environments with applications to a cognitively-controlled hearing aid," in *Neural Netw.*, vol. 140, pp. 136–147, 2021.

[12] T. Green et al., "Speech recognition with a hearing-aid processing scheme combining beamforming with mask-informed speech enhancement," *Trends Hear.*, vol. 26, pp. 1–16, 2022. [Online]. Available: https://journals.sagepub.com/doi/10.1177/23312165211068629

[13] T. Gajecki and W. Nogueira, "Deep Latent Fusion Layers for Binaural Speech Enhancement," in *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 3127–3138, 2023.

[14] V. Tokala, E. Grinstein, M. Brookes, S. Doclo, J. Jensen, and P. A. Naylor, "Binaural speech enhancement using deep complex convolutional transformer networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2024, pp. 681–685. [Online]. Available: https://ieeexplore.ieee.org/document/10447090/?arnumber=10447090

[15] M. Tammen and S. Doclo, "Deep multi-frame MVDR filtering for binaural noise reduction," in *Proc. IEEE Int. Workshop Acoust. Signal Enhancement*, 2022, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/9914742/

[16] W. Mack and E. A. P. Habets, "Deep filtering: Signal extraction and reconstruction using complex time-frequency filters," *IEEE Signal Process. Lett.*, vol. 27, pp. 61–65, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8911434/

[17] N. L. Westhausen and B. T. Meyer, "Low bit rate binaural link for improved ultra low-latency low-complexity multichannel speech enhancement in hearing aids," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, 2023, pp. 1–5.

[18] T. Nakatani et al., "DNN-supported mask-based convolutional beamforming for simultaneous denoising, dereverberation, and source separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2020, pp. 6399–6403.

[19] Z. Ni et al., "WPD : An improved neural beamformer for simultaneous speech separation and dereverberation," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, 2021, pp. 817–824.

[20] Z. Zhang et al., "Multi-Channel Multi-Frame ADL-MVDR for Target Speech Separation," in *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3526–3540, 2021.

[21] Y. Xu, Z. Zhang, M. Yu, S.-X. Zhang, and D. Yu, "Generalized spatio-temporal RNN beamformer for target speech separation," in *Proc. Interspeech*, 2021, pp. 3076–3080 [Online]. Available: https://www.isca-speech.org/archive/interspeech_2021/xu21i_interspeech.html

[22] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Integrating full- and sub-band modeling for speech separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 3221–3236, 2023. [Online]. Available: https://doi.org/10.1109/TASLP.2023.3304482

[23] M. Tammen and S. Doclo, "Parameter Estimation Procedures for Deep Multi-Frame MVDR Filtering for Single-Microphone Speech Enhancement," *IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 31, pp. 3237–3248, 2021.

[24] J. Benesty, J. Chen, and E. A. P. Habets, *Speech Enhancement in the STFT Domain*. Heidelberg, Germany: Springer, 2011.

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

[26] J. Casebeer, J. Donley, D. Wong, B. Xu, and A. Kumar, "NICE-Beam: Neural integrated covariance estimators for time-varying beamformers," 2021, *arXiv:2112.04613*.

[27] C. K. A. Reddy et al., "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Proc. Interspeech*, 2020, pp. 2492–2496 [Online]. Available: https://www.isca-speech.org/archive/interspeech_2020/reddy20_interspeech.html

[28] C. K. A. Reddy et al., "INTERSPEECH 2021 deep noise suppression challenge," in *Proc. Interspeech*, 2021, pp. 2796–2800. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2021/reddy21_interspeech.html

[29] S. Graetzer et al., "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in *Proc. Interspeech*, 2021, pp. 686–690. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2021/graetzer21_interspeech.html

[30] B. Cornelis, S. Doclo, T. Van dan Bogaert, M. Moonen, and J. Wouters, "Theoretical analysis of binaural multimicrophone noise reduction techniques," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 2, pp. 342–355, Feb., 2010. [Online]. Available: https://ieeexplore.ieee.org/document/5173569/

[31] E. A. P. Habets, J. Benesty, and J. Chen, "Multi-microphone noise reduction using interchannel and interframe correlations," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2012, pp. 305–308.

[32] J. Pak and J. W. Shin, "Sound localization based on phase difference enhancement using deep neural networks," in *IEEE Trans. Audio Speech Lang. Process.*, vol. 27, no. 8, pp. 1335–1345, Aug., 2019.

[33] J.-M. Lemercier, J. Thiemann, R. Koning, and T. Gerkmann, "A neural network-supported two-stage algorithm for lightweight dereverberation on hearing devices," *EURASIP J. Audio, Speech, and Music Process.*, vol. 2023, no. 1, 2023, Art. no. 18. [Online]. Available: https://asmp-eurasipjournals.springeropen.com/articles/10.1186/s13636-023-00285-8

[34] G. H. Golub and C. F. Van Loan, Matrix Computations Baltimore, MD, USA: Johns Hopkins Univ. Press, 2013.

[35] Z.-Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR," in *IEEE Trans. Audio Speech Lang. Process.*, vol. 28, pp. 1778–1787, 2020.

[36] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2017, pp. 1–18 [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[37] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, Art. no. 721.

[38] M. Herdin, N. Czink, H. Ozcelik, and E. Bonek, "Correlation matrix distance, a meaningful measure for evaluation of non-stationary MIMO channels," in *Proc. IEEE 61st Veh. Technol. Conf. (VTC)*, 2005, pp. 136–140. [Online]. Available: https://ieeexplore.ieee.org/document/1543265/

[39] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) - A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2001, pp. 749–752.

[40] J. M. Kates and K. H. Arehart, "The hearing-aid speech quality index (HASQI)," in *J. Audio Eng. Soc.*, vol. 58, no. 5, pp. 363–381, 2010.

[41] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI) version 2," *Speech Commun.*, vol. 131, pp. 35–46, 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167639320300431

Marvin Tammen (Graduate Student Member, IEEE) received the B.Eng. degree in 2015 and the M.Sc. degree in 2018 from the University of Oldenburg, Oldenburg, Germany, where he is currently working toward the Dr.-Ing. degree with Signal Processing Group. In 2021 and 2023, he was a Research Intern with Facebook Reality Labs and NTT Communication Science Laboratories, respectively. His research interests include fusing model-based and deep learning-based approaches for speech enhancement.

Simon Doclo (Senior Member, IEEE) received the M.Sc. degree in electrical engineering and the Ph.D. degree in applied sciences from KU Leuven, Leuven, Belgium, in 1997 and 2003, respectively. From 2003 to 2007 he was a Postdoctoral Fellow with KU Leuven and McMaster University, Hamilton, ON, Canada. From 2007 to 2009 he was a Principal Scientist with NXP Semiconductors, Leuven. Since 2009, he has been a Full Professor with the University of Oldenburg, and Scientific Advisor for the Branch Hearing, Speech and Audio Technology HSA of the Fraunhofer Institute for Digital Media Technology IDMT, Ilmenau, Germany. His research interests include signal processing for acoustical and biomedical applications, more specifically microphone array processing, speech enhancement, active noise control, acoustic sensor networks and hearing aid processing. Dr. Doclo was the recipient of the several best paper awards (International Workshop on Acoustic Echo and Noise Control 2001, EURASIP Signal Processing 2003, IEEE Signal Processing Society 2008, VDE Information Technology Society 2019). He was a Member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing and the EURASIP Technical Area Committee on Acoustic, Speech and Music Signal Processing and a Member of the EAA Technical Committee on Audio Signal Processing. Since 2021, he has been aSenior Area Editor of IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING.