Speech Enhancement Using Masking for Binaural Reproduction of Ambisonics Signals

Moti Lugasi, *Student Member, IEEE*, and Boaz Rafaely, *Senior Member, IEEE*

***Abstract*—Speech enhancement in a single channel has been well studied in the literature in applications such as speech communication systems. However, in emerging applications such as virtual reality and spatial audio, in addition to attenuating undesired signals, the ability to preserve the spatial information of the desired signal captured in a noisy environment is of great importance. Nevertheless, there are only a few studies in the literature that propose solutions to this challenge. Most of these studies present solutions that attenuate the undesired signals, while preserving only limited spatial information regarding the desired signal, such as the direction of arrival (DOA). Methods that preserve complete spatial information have only recently been suggested, and have not been studied comprehensively. In this paper, two such methods based on time-frequency masking are investigated with the aim of attenuating the undesired signal, while preserving the spatial components of the desired signal. The first is referred to as spatial masking and is based on masking in the plane wave density (PWD) domain, and the second on masking in the spherical harmonics (SH) domain. The two methods are compared with a reference method, based on beamforming followed by single-channel time-frequency masking. Objective analysis and two listening tests were conducted in order to evaluate the performance of these methods for speech enhancement. It was shown that the spatial masking based method better preserves the desired component of the sound field, while the performance of the SH based method more strongly depends on the sources' distances. On the other hand, the SH based method better preserves the DOA of the residual noise, while the DOA of the residual noise under the spatial masking based method is strongly affected by the undesired signal.**

***Index Terms*—Speech enhancement, Wiener masking, spatial masking, plane wave decomposition, spherical arrays, binaural reproduction, noise reduction.**

I. Introduction

EVERYDAY sounds received at the listener's ears carry spatial information about the distance and direction of the sound source relative to the listener, in addition to temporal information from the source signal. This spatial information is also important when reproducing natural sound in a variety of applications that incorporate spatial audio, which in recent years has become increasingly popular. Virtual reality is one such application, featuring in education, gaming and entertainment, architectural design, and more [1]–[4]. Other applications based on spatial audio include speech communication [5], [6], and aids for the hearing impaired [7]–[9].

Spatial audio signals can be generated by artificially creating the audio signals, or by capturing the signals from the real world [10]. The latter is important in applications such as recording music events; communication in video conferencing meetings; and as a companion to 3D video capture. Although significant progress has already been made in spatial audio, there still remain many challenges in the recording and reproduction of acoustic scenes. One of these challenges arises from the limitations of the audio recording systems. Specifically, the quality of reproduction may be limited by the spatial information delivered by practical recording systems [11], [12]. Spherical microphone arrays, in particular, have been studied in this context due to the spherical harmonics (SH) processing, leading to the well established Ambisonics format [13]–[16]. Nevertheless, even with high quality recording systems, the sound field in the real world may include, in addition to desired components such as speech or music, also undesired components such as noise or other interferences. The challenge, in this case, is to attenuate the undesired components without distorting the spatial information in the desired components. While some studies have offered limited solutions (discussed below), the problem of enhancing spatial audio signals in general, and speech signals in particular, remains, to a large extent, open.

One recently published approach for speech enhancement in a single channel is based on deep neural networks (DNNs), which have been found to be very useful in preserving monaural information [17], [18]. Another approach for spatial signal enhancement involves standard array signal processing methods, which may be applied to cancel undesired signal components and produce a single output channel [19], [20]. While useful for signal enhancement, these approaches do not preserve the spatial information that is essential for sound reproduction. Additional approaches, for hearing aids applications, for example, use a binaural beamformer in order to attenuate the undesired signals and preserve the desired signals without distortion [9], [21]. In order to attenuate the undesired signals, these methods, which are based on beamforming, assume spatial separation between the desired and the undesired signals. This assumption may not hold in case of a reverberant environment. To overcome this limitation, a more advanced method [22] uses beamforming and applies time-frequency masking on the binaural signals. With this approach, the undesired signal components arriving from the desired signal's direction can be attenuated. Nevertheless, due to beamformer limitations, the number of constraints in

Manuscript received July 7, 2019; revised January 22, 2020 and May 14, 2020; accepted May 22, 2020. Date of publication May 28, 2020; date of current version June 18, 2020. This work was supported by Facebook Reality Labs. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jun Du. *(Corresponding author: Moti Lugasi.)*

The authors are with the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel (e-mail: motilu@post.bgu.ac.il; br@bgu.ac.il).

Digital Object Identifier 10.1109/TASLP.2020.2998294

this approach cannot exceed the number of the system's degrees of freedom. Thus, these methods can only preserve the spatial information at selected directions.

To preserve more of the spatial information in the desired signal, the authors in [23] present a simple formulation of the binaural signals as a function of the beamformer's coefficients in the SH domain. However, it has been shown in [24] that there is a trade-off between the noise reduction and the preservation of the spatial information of the desired signal at the output. Preservation of spatial information is therefore still limited, in particular when substantial reduction of the undesired signal is required.

The authors in [25] propose an algorithm that preserves the entire spatial information of the desired signal by using a beamformer to extract the desired source signal, and then estimating the transfer function from the source to the array, leading to accurate reproduction of spatial information of the desired signal. However, because this method assumes one dominant source at the beamformer's look direction, estimation of the source signal may degrade when the distance between the microphone array and the source is larger than the critical distance. Moreover, in the case where the undesired signal is diffuse noise, the noise part arriving from the desired source direction is not attenuated. As a result, the spatial information of the desired signal is degraded.

To overcome the limitations of the aforementioned methods, approaches combining spatial masking, i.e. masking applied in the plane wave density (PWD) domain, and time-frequency masking were developed. In [26] the authors presented a method based on Wiener masking in the spatial domain. For this method, the mask is calculated using the ratio between the spatially-localized spherical Fourier transform (SLSFT) of the desired and the undesired signal components, and its performance is compared to that of another method, where the mask is calculated using the ratio between the SH coefficients of the desired and the undesired signals, and analysed for stationary signals. Possible extension of these methods to non stationary signals, such as speech, could involve time-frequency masking as well. In a recent work, two approaches based on time-frequency filtering were presented [27]. The first method requires knowledge of the steering vectors of the desired sources and the second approach uses a special direction-preserving form of the filter. For both approaches, the parametric multichannel Wiener filter (PMWF) [28] is employed. This filter is formulated in the SH domain to provide an Ambisonics representation of the desired signals. As discussed in [27], the performance of the first approach may be very limited in highly reverberant sound environments. The second approach is further investigated here.

In this paper, methods that aim to preserve the entire desired sound field, using time-frequency masking were investigated. These include: (i) the SH mask approach from [26], which is extended to the time-frequency domain; (ii) the second approach outlined in [27], which is based on a spatial time-frequency mask; and (iii) a third method, chosen as a low-end reference and motivated by [22], based on a single beamformer with a time-frequency mask. The focus in the paper on masking-based methods stems from the simplicity of application of this approach. Other methods mentioned in this introduction were not investigated as they do not claim to preserve the entire desired sound field, and thus do not ally with the focus. All three approaches are presented and evaluated for binaural reproduction in a reverberant environment. These approaches are formulated as Wiener masking in the time-frequency domain applied to the observed signal (i.e. the noisy signal) in the spatial and the SH domains. The performance of these methods was evaluated under ideal conditions, i.e. the Wiener mask was computed given the oracle information of the instantaneous signal-to-noise ratio (SNR) at every time-frequency bin. For objective analysis, measures of the SNR gain and the signal to distortion ratio (SDR) of the binaural signals after applying each method are employed. Inter-aural level difference (ILD) and inter-aural cross-correlation time (IACC_t) time of the residual noise are also computed for the binaural signals, providing additional performance measures. Finally, two listening tests were conducted to provide further insight into the performance of these methods for binaural reproduction.

The paper presents the following contributions: (i) a comparison of recently proposed methods for the enhancement of spatial audio speech signals, focusing on methods that do not require a priori spatial information, but, rather, rely on standard methods for estimating SNR; (ii) specifically showing the advantage of masking in the spatial domain relative to masking directly in the SH (Ambisonics) domain, with the latter showing higher sensitivity to the sources' distances from the microphone array; (iii) validation of performance through listening tests, complementing the objective performance analysis (to date, subjective evaluation has not been presented in the recent papers proposing the various approaches); (iv) recommendations for the enhancement of Ambisonics signals using masking in the spatial domain, for signals with a dominant direct component, supporting partial results of recent studies; (v) recommendations for the enhancement of Ambisonics signals using masking in the SH domain in highly reverberant environments with distant speakers, due to the direct processing in the SH domain; (vi) recommendations for the enhancement of Ambisonics signals using masking in the SH domain in the case where the spatial cues of the residual noise need to be preserved.

This paper is structured as follows. In Section II the system model and commonly used notations are described. In Section III three Wiener masking methods are presented, and in Section IV binaural reproduction formulations are derived using the SH estimators of the desired signal from Section III. In Section V objective measures are defined, and in Section VI objective analysis is conducted to evaluate the performance of the masking methods from Section III. In Sections VII and VIII two listening tests are performed to further study and validate the properties of the methods. Section IX presents the conclusions and suggestions for future investigation of the presented masking methods.

II. System Model

This section presents the system model that will be used throughout this work. We consider a spherical array of arbitrary configuration located at the origin of the Cartesian coordinates, and composed of J omnidirectional sensors. The j-th element of this array is located at $\mathbf{r}_j = (r\cos\phi_j \sin\theta_j, r\sin\theta_j \sin\phi_j, r\cos\theta_j)^T$. The elevation

angle θ is measured downwards from the positive z-axis, the azimuth angle is measured from the positive x-axis towards the positive y-axis, and r is the Euclidean distance of the j-th element from the origin. The array is assumed to be positioned in a reverberant sound field. According to the image method [29], the sound pressure in a reverberant environment, which is generated by a single source in the far field, can be modeled as a sum of L significant plane waves which are generated by L image sources in free-field. Assuming that $\Psi_l = (\theta_l, \phi_l)$ is the direction of arrival (DOA) of the l-th sources, the sound pressure at the sensors can be expressed as [19]:

$$\mathbf{p}(k) = \sum_{l=1}^{L} \mathbf{v}(k, \Psi_l) s_l(k), \tag{1}$$

where $k = \frac{2\pi f}{c}$ is the wave number, f is the frequency and c is the speed of sound. $\mathbf{p}(k)$ is defined to be the vector of sound pressure at the array's sensors at wave number k, $\mathbf{p}(k) = [p_1(k), p_2(k), \ldots, p_J(k)]^T$. $\mathbf{v}(k, \Psi_l)$ is the steering vector at direction Ψ_l and $s_l(k)$ is the complex amplitude of the l-th source signal. The matrix form of Eq. (1) is given by:

$$\mathbf{p}(k) = \mathbf{V}(k, \mathbf{\Psi})\mathbf{s}(k), \tag{2}$$

where

$$\mathbf{V}(k, \mathbf{\Psi}) = [\mathbf{v}(k, \Psi_1), \mathbf{v}(k, \Psi_2), \ldots, \mathbf{v}(k, \Psi_L)]^T \tag{3}$$

is a $J \times L$ steering matrix,

$$\mathbf{s}(k) = [s_1(k), s_2(k), \ldots, s_L(k)]^T \tag{4}$$

is the vector of source signals at wave number k, and vector $\mathbf{\Psi} = [\Psi_1, \Psi_2, \ldots, \Psi_L]^T$ is the vector of the DOAs.

In the case of spherical arrays, the inverse spherical Fourier transform of the sound pressure $p(k, r, \Omega)$ at frequency k and angle $\Omega = (\theta, \phi)$ on the surface of a sphere with radius r is defined by:

$$p(k, r, \Omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} p_{nm}(k, r) Y_n^m(\Omega), \tag{5}$$

where $Y_n^m(\Omega)$ denotes the SH functions of order n and degree m [30] and p_{nm} are the coefficients of the spherical Fourier transform. As explained in [30], the coefficients p_{nm} diminish for $n > kr$ and can therefore be neglected. Hence, Eq. (5) can be approximated for an appropriate finite order N:

$$p(k, r, \Omega) \cong \sum_{n=0}^{N} \sum_{m=-n}^{n} p_{nm}(k, r) Y_n^m(\Omega). \tag{6}$$

Consequently, the steering matrix $\mathbf{V}(k, \mathbf{\Psi})$ can also be approximated for a finite order N and decomposed as follows (for more discussion of this decomposition the reader is referred to [31]):

$$\mathbf{V}(k, \mathbf{\Psi}) = \mathbf{Y}(\mathbf{\Omega})\mathbf{B}(kr)\mathbf{Y}^H(\mathbf{\Psi}), \tag{7}$$

where $\mathbf{Y}(\mathbf{\Psi})$ is an $L \times (N+1)^2$ matrix of SH, defined as:

$$\mathbf{Y}(\mathbf{\Psi}) = \begin{bmatrix} \mathbf{y^T}(\mathbf{\Psi_1}) \\ \mathbf{y}^T(\Psi_2) \\ \vdots \\ \mathbf{y}^T(\Psi_L) \end{bmatrix}, \tag{8}$$

where $\mathbf{y}(\Psi_l) = [Y_0^0(\Psi_l), Y_1^{-1}(\Psi_l) Y_1^0(\Psi_l), \ldots, Y_N^N(\Psi_l)]^T$. Matrix $\mathbf{B}(kr) = \text{diag}(b_0, b_1, b_1, b_1, \ldots, b_N)$ is an $(N+1)^2 \times (N+1)^2$ diagonal matrix with elements $\{b_n(kr)\}_{n=0}^N$ representing radial functions that depend on the wave number and distance from the origin [30]. Matrix $\mathbf{Y}(\mathbf{\Omega})$, of dimensions $J \times (N+1)^2$, is defined in a manner similar to Eq. (8), where $\mathbf{\Omega} = [\Omega_1, \Omega_2, \ldots, \Omega_J]$ is a $J \times 1$ vector of the sensors' locations on the sphere. N is the highest SH order of the representation, which is typically chosen to satisfy $kr < N$ to avoid excessive truncation errors [12], [30]. Multiplying Eq. (2) by the pseudo-inverse of $\mathbf{Y}(\mathbf{\Omega})\mathbf{B}(kr)$, which is defined as $(\mathbf{YB})^\dagger = [(\mathbf{YB})^H(\mathbf{YB})]^{-1}(\mathbf{YB})^H$, and assuming the number of sensors J is larger than the number of SH coefficients $(N+1)^2$, leads to the SH decomposition [32]:

$$\mathbf{a}_{nm}(k) = \mathbf{Y}^H(\mathbf{\Psi})\mathbf{s}(k), \tag{9}$$

where $\mathbf{a}_{nm}(k) = [a_{00}(k), a_{1(-1)}(k), a_{10}(k), \ldots, a_{NN}(k)]^T$ is the $(N+1)^2 \times 1$ vector of the PWD coefficients in the SH domain [31], also denoted high-order Ambisonics signals [33]. By using the short-time Fourier transform (STFT), Eq. (9) can be presented in the time-frequency domain as:

$$\mathbf{a}_{nm}(\tau, \nu) = \mathbf{Y}^H(\mathbf{\Psi})\mathbf{s}(\tau, \nu), \tag{10}$$

where τ is the time index and ν is the frequency index. Finally, the PWD function is calculated by using the inverse spherical Fourier transform [30] and Eq. (10):

$$\mathbf{a}(\mathbf{\Phi}, \tau, \nu) = \mathbf{Y}(\mathbf{\Phi})\mathbf{a}_{nm}(\tau, \nu), \tag{11}$$

where $\mathbf{a}(\mathbf{\Phi}, \tau, \nu)$ is the $Q \times 1$ vector of the PWD function at $\mathbf{\Phi} = [\Phi_1, \Phi_2, \ldots, \Phi_Q]$ arbitrary directions.

III. Wiener Masking Methods

In real scenarios, the sound pressure may be composed of desired components and undesired components. This section presents three different methods based on Wiener masking that aim to attenuate the undesired components, while preserving the desired components of the signal. It has been shown that the Wiener mask is a very effective method for signal enhancement in single and multiple channel systems where the signal is contaminated by noise [34]. By using a Wiener mask, the intelligibility and the quality of noisy speech signals can be improved [35], [36]. Hence, the Wiener mask is widely used in speech communication systems [34], [36]. The Wiener mask is generally applied to the noisy signal in the time-frequency domain, either separately to each microphone in an array, or to the output of an array beamformer. Even though the multi channel Wiener filter has been well studied, its most established form yields a single-channel output, while its formulation with multiple channel outputs, has not been studied extensively. In

the proposed methods, the Wiener mask is applied to the signal, which is presented in the spatial and SH domains, as well as in the time-frequency domain. While some works have been published that pursued this research direction [22], [26], [27], these methods are not well established, so we will expand here. As presented in the introduction, various approaches for the enhancement of spatial audio signals have been proposed, while this paper focuses on approaches that require no, or very limited, a priori information on the signals, and only apply standard methods for SNR estimation. These approaches could therefore be useful for the direct enhancement of Ambisonics signals without the need for sound field parametrisation. For the methods presented below, the instantaneous SNR is defined using oracle information, such that the desired and the undesired components are separately available in the time-frequency domain. In practice, however, methods for SNR estimation must be applied [37], [38].

The sound pressure is represented as composed of desired and undesired components, following the notation in Eq. (1) as:

$$\mathbf{p}(k) = \sum_{l=1}^{L} \mathbf{v}(k, \Psi_l^d) s_l(k) + \sum_{i=1}^{I} \mathbf{v}(k, \Psi_i^u) n_i(k), \tag{12}$$

where $s_l(k)$ is the complex amplitude of the l-th desired source signal and $n_i(k)$ is the complex amplitude of the i-th undesired source signal. By following the same derivation as in Eqs. (1) to (10), the PWD coefficients in the STFT domain can be written as:

$$\begin{aligned}\mathbf{a}_{nm}(\tau,\nu) &= \mathbf{Y}^H(\mathbf{\Psi_d})\mathbf{s}(\tau,\nu) + \mathbf{Y}^H(\mathbf{\Psi_u})\mathbf{n}(\tau,\nu) \\ &= \mathbf{a}^d_{nm}(\tau,\nu) + \mathbf{a}^u_{nm}(\tau,\nu),\end{aligned} \tag{13}$$

where $\mathbf{a}^d_{nm}(\tau,\nu) = \mathbf{Y}^H(\mathbf{\Psi_d})\mathbf{s}(\tau,\nu)$ and $\mathbf{a}^u_{nm}(\tau,\nu) = \mathbf{Y}^H(\mathbf{\Psi_u})\mathbf{n}(\tau,\nu)$, $\mathbf{\Psi_d} = [\Psi_1^d, \Psi_2^d, \ldots, \Psi_L^d]^T$ is the DOA vector of the desired components, $\mathbf{\Psi_u} = [\Psi_1^u, \Psi_2^u, \ldots, \Psi_I^u]^T$ is the DOA vector of the undesired components and $\mathbf{s}(\tau,\nu)$ and $\mathbf{n}(\tau,\nu)$ are the vectors of the desired and undesired source signals in the STFT domain, with lengths L and I, respectively. The PWD function can be computed by substituting Eq. (13) in Eq. (11):

$$\begin{aligned}\mathbf{a}(\mathbf{\Phi},\tau,\nu) &= \mathbf{Y}(\mathbf{\Phi})\mathbf{Y}^H(\mathbf{\Psi_d})\mathbf{s}(\tau,\nu) + \mathbf{Y}(\mathbf{\Phi})\mathbf{Y}^H(\mathbf{\Psi_u})\mathbf{n}(\tau,\nu) \\ &= \mathbf{a}_d(\mathbf{\Phi},\tau,\nu) + \mathbf{a}_u(\mathbf{\Phi},\tau,\nu),\end{aligned} \tag{14}$$

where $\mathbf{a}_d(\mathbf{\Phi},\tau,\nu) = \mathbf{Y}(\mathbf{\Phi})\mathbf{Y}^H(\mathbf{\Psi_d})\mathbf{s}(\tau,\nu)$ and $\mathbf{a}_u(\mathbf{\Phi},\tau,\nu) = \mathbf{Y}(\mathbf{\Phi})\mathbf{Y}^H(\mathbf{\Psi_u})\mathbf{n}(\tau,\nu)$. Eqs. (13) and (14) are the representations of the signal in the space and SH domains, respectively. Masking is formulated in the following sections based on these representations.

A. Time-Frequency-Spherical Harmonics Mask (TFSH Mask)

The TFSH mask is applied to the coefficients of the PWD function in the SH domain. This mask was suggested in [26] for stationary signals, while here it is extended to the time-frequency domain. In this case the Wiener mask is defined as:

$$M(n,m,\tau,\nu) = \frac{SNR(n,m,\tau,\nu)}{SNR(n,m,\tau,\nu)+1}, \tag{15}$$

where

$$SNR(n,m,\tau,\nu) = \frac{E[|a^d_{nm}(\tau,\nu)|^2]}{E[|a^u_{nm}(\tau,\nu)|^2]}, \tag{16}$$

and $a^d_{nm}(\tau,\nu)$ and $a^u_{nm}(\tau,\nu)$ are the nm-th element of the vectors $\mathbf{a}^d_{nm}(\tau,\nu)$ and $\mathbf{a}^u_{nm}(\tau,\nu)$, respectively, and $E[\cdot]$ denotes expectation. The instantaneous evaluation of the SNR in Eq. (16) is given by:

$$SNR(n,m,\tau,\nu) = \frac{|a^d_{nm}(\tau,\nu)|^2}{|a^u_{nm}(\tau,\nu)|^2}. \tag{17}$$

The estimator $\hat{\mathbf{a}}^d_{nm}(\tau,\nu)$ of $\mathbf{a}^d_{nm}(\tau,\nu)$ from $\mathbf{a}_{nm}(\tau,\nu)$ at specific time-frequency bins is given by:

$$\hat{\mathbf{a}}^d_{nm}(\tau,\nu) = \mathbf{M}(n,m,\tau,\nu)\mathbf{a}_{nm}(\tau,\nu), \tag{18}$$

where $\mathbf{M}(n,m,\tau,\nu)$ is a $(N+1)^2 \times (N+1)^2$ diagonal matrix defined by using Eq. (15) as:

$$\begin{aligned}&\mathbf{M}(n,m,\tau,\nu) \\ &= \mathrm{diag}(M(0,0,\tau,\nu), M(1,(-1),\tau,\nu), \ldots, M(N,N,\tau,\nu)).\end{aligned} \tag{19}$$

B. Time-Frequency-Space Mask (TFS Mask)

The TFS mask is applied to the PWD function. In this case the Wiener mask is defined as:

$$M(\Phi_q,\tau,\nu) = \frac{SNR(\Phi_q,\tau,\nu)}{SNR(\Phi_q,\tau,\nu)+1}, \tag{20}$$

where

$$SNR(\Phi_q,\tau,\nu) = \frac{E[|a_d(\Phi_q,\tau,\nu)|^2]}{E[|a_u(\Phi_q,\tau,\nu)|^2]}, \tag{21}$$

and the q-th element of vector $\mathbf{a}_d(\mathbf{\Phi},\tau,\nu)$ and vector $\mathbf{a}_u(\mathbf{\Phi},\tau,\nu)$ is defined to be $a_d(\Phi_q,\tau,\nu)$ and $a_u(\Phi_q,\tau,\nu)$, respectively. The instantaneous evaluation of the SNR in Eq. (21) is given by:

$$SNR(\Phi_q,\tau,\nu) = \frac{|a_d(\Phi_q,\tau,\nu)|^2}{|a_u(\Phi_q,\tau,\nu)|^2}. \tag{22}$$

A diagonal $Q \times Q$ matrix of the Wiener mask in Eq. (20), which is calculated for $q = 1, \ldots, Q$, can be defined as:

$$\begin{aligned}\widetilde{\mathbf{M}}(\mathbf{\Phi},\tau,\nu) = \mathrm{diag}(&M(\Phi_1,\tau,\nu), M(\Phi_2,\tau,\nu), \ldots, \\ &M(\Phi_Q,\tau,\nu)).\end{aligned} \tag{23}$$

The estimator $\hat{\mathbf{a}}_d(\mathbf{\Phi},\tau,\nu)$ of $\mathbf{a}_d(\mathbf{\Phi},\tau,\nu)$ from $\mathbf{a}(\mathbf{\Phi},\tau,\nu)$ in a specific time-frequency bin is given by:

$$\hat{\mathbf{a}}_d(\mathbf{\Phi},\tau,\nu) = \widetilde{\mathbf{M}}(\mathbf{\Phi},\tau,\nu)\mathbf{a}(\mathbf{\Phi},\tau,\nu). \tag{24}$$

By using Eq. (11), Eq. (24) can be rewritten as:

$$\hat{\mathbf{a}}^d_{nm}(\tau,\nu) = \mathbf{Y}^\dagger(\mathbf{\Phi})\widetilde{\mathbf{M}}(\mathbf{\Phi},\tau,\nu)\mathbf{Y}(\mathbf{\Phi})\mathbf{a}_{nm}(\tau,\nu), \tag{25}$$

where $\mathbf{Y}^\dagger(\mathbf{\Phi}) = [\mathbf{Y}^H(\mathbf{\Phi})\mathbf{Y}(\mathbf{\Phi})]^{-1}\mathbf{Y}^H(\mathbf{\Phi})$. By denoting $\mathbf{M}(\mathbf{\Phi},\tau,\nu) = \mathbf{Y}^\dagger(\mathbf{\Phi})\widetilde{\mathbf{M}}(\mathbf{\Phi},\tau,\nu)\mathbf{Y}(\mathbf{\Phi})$, Eq. (25) can be rewritten in a simpler form as:

$$\hat{\mathbf{a}}^d_{nm}(\tau,\nu) = \mathbf{M}(\mathbf{\Phi},\tau,\nu)\mathbf{a}_{nm}(\tau,\nu). \tag{26}$$

C. Beamforming Followed by Masking

Motivated by [22], this method uses beamforming and a time-frequency mask in order to preserve the desired signal at the source direction only. This method is suggested as a lower-end reference to the TFSH mask and the TFS mask.

By applying a beamformer in the SH domain, the array output is given by:

$$z(\tau,\nu) = \mathbf{w}_{nm}^H \mathbf{a}_{nm}(\tau,\nu), \tag{27}$$

where $\mathbf{a}_{nm}(\tau,\nu)$ is defined in Eq. (13) and $\mathbf{w}_{nm}$ is the SH coefficients vector of an arbitrary beamformer [30]. Let Ω_s be the DOA of the source, which for this case is assumed to be known. In the case of a maximum directivity beamformer with look direction Ω_s, Eq. (27) can be rewritten as:

$$z(\tau,\nu) = \mathbf{y}^T(\Omega_s)\mathbf{a}_{nm}(\tau,\nu), \tag{28}$$

where $\mathbf{y}(\Omega_s)$ is defined in Eq. (8). By using Eq. (13), Eq. (28) can be rewritten as:

$$\begin{aligned} z(\tau,\nu) &= \mathbf{y}^T(\Omega_s)\mathbf{a}_{nm}^d(\tau,\nu) + \mathbf{y}^T(\Omega_s)\mathbf{a}_{nm}^u(\tau,\nu) \\ &= z_d(\tau,\nu) + z_u(\tau,\nu), \end{aligned} \tag{29}$$

where $z_d(\tau,\nu) = \mathbf{y}^T(\Omega_s)\mathbf{a}_{nm}^d(\tau,\nu)$ and $z_u(\tau,\nu) = \mathbf{y}^T(\Omega_s)\mathbf{a}_{nm}^u(\tau,\nu)$ are the desired and the undesired signals as filtered by the beamformer with look direction Ω_s, respectively. In order to extract the desired signal at direction Ω_s, a Wiener mask is applied to $z(\tau,\nu)$ in the time-frequency domain. In this case the Wiener mask is defined as:

$$M(\tau,\nu) = \frac{SNR(\tau,\nu)}{SNR(\tau,\nu)+1}, \tag{30}$$

where

$$SNR(\tau,\nu) = \frac{E[|z_d(\tau,\nu)|^2]}{E[|z_u(\tau,\nu)|^2]}. \tag{31}$$

The instantaneous evaluation of the SNR in Eq. (31) is given by:

$$SNR(\tau,\nu) = \frac{|z_d(\tau,\nu)|^2}{|z_u(\tau,\nu)|^2}. \tag{32}$$

The estimator $\hat{z}_d(\tau,\nu)$ of $z_d(\tau,\nu)$ from $z(\tau,\nu)$ is given by:

$$\hat{z}_d(\tau,\nu) = M(\tau,\nu)z(\tau,\nu). \tag{33}$$

IV. Binaural Reproduction

Binaural reproduction is the final stage of the process, once the desired signal has been calculated using each of the methods described in the previous section. As shown in [39] the sound pressure at the right ear, $P_r(k)$, and the left ear, $P_r(k)$, at frequency k, is calculated using:

$$P_{r,l}(k) = \int_{\Omega\in S^2} a(\Omega,k)H_{r,l}(\Omega,k)d\Omega, \tag{34}$$

where $H_{r,l}(\Omega,k)$ is the head related transfer function (HRTF) of the right ear $H_r(\Omega,k)$ and the left ear $H_l(\Omega,k)$ at frequency k and directions $\Omega \in S^2$, and $a(\Omega,k)$ is the PWD function at frequency k and directions $\Omega \in S^2$. Eq. (34) can be approximated by a finite sum of SH coefficients [40]:

$$P_{r,l}(k) \cong \sum_{n=0}^{N}\sum_{m=-n}^{n} [\tilde{a}_{nm}(k)]^* H_{nm}^{r,l}(k), \tag{35}$$

where $\tilde{a}_{nm}(k) = (-1)^m[a_{n(-m)}(k)]^*$ is the representation of $a^*(\Omega,k)$ in the SH domain. For TFS and TFSH masks, the binaural signal is reproduced by using the estimator of the desired signal represented in the frequency domain ($\hat{\mathbf{a}}_{nm}^d(k)$) and Eq. (35):

$$P_{r,l}^M(k) \cong \sum_{n=0}^{N}\sum_{m=-n}^{n} [\tilde{\hat{a}}_{nm}^d(k)]^* H_{nm}^{r,l}(k), \tag{36}$$

where $\hat{a}_{nm}^d(k)$ is the element of the vector $\hat{\mathbf{a}}_{nm}^d(k)$ with order n and degree m. In the same manner, $P_{r,l}^d(k)$ and $P_{r,l}^u(k)$ are the binaural signals computed using $\mathbf{a}_{nm}^d(k)$ and $\mathbf{a}_{nm}^u(k)$, respectively.

For the beamforming method, due to the single channel output, as in Eq. (28), only the HRTF in direction Ω_s is used:

$$P_{r,l}^M(k) = z(k)H_{r,l}(\Omega_s,k), \tag{37}$$

where $z(k)$ is the representation of Eq. (33) in the frequency domain.

V. Objective Measures of Performance

In this section objective measures are formulated in order to evaluate the performance of the proposed methods. The input signal is composed of desired and undesired components, as shown in Eq. (13). The signal to noise ratio between the desired and undesired components is defined as:

$$SNR_{in} = 10\log_{10}\left(\frac{\sum_{t=1}^{T}||\mathbf{a}_{nm}^d(t)||^2}{\sum_{t=1}^{T}||\mathbf{a}_{nm}^u(t)||^2}\right), \tag{38}$$

where SNR_{in} reflects the ratio between the desired signal's energy and the undesired signal's energy in the SH domain, $\mathbf{a}_{nm}^d(t)$ and $\mathbf{a}_{nm}^u(t)$ are the inverse STFTs of the vectors defined in Eq. (13) and $||\cdot||$ denotes the Euclidean norm. The processing methods presented in Sec. III and the binaural reproduction formulation presented in Sec. IV can now be applied to the input signal composed of these two separate components: desired and undesired. In this case, the signals at the ears produced through binaural reproduction and estimation of the desired signal are given by:

$$P_{r,l}^M(k) = P_{r,l}^{Md}(k) + P_{r,l}^{Mu}(k), \tag{39}$$

where $P_{r,l}^{Md}(k)$ and $P_{r,l}^{Mu}(k)$ are the desired and undesired components of the binaural signals, respectively, after applying the processing methods presented in Section III. Eq. (39) replaces Eqs. (36) and (37) after formulating Eqs. (18), (26) and (33) as a superposition of desired and undesired components.

Next, the improvement in the SNR of the binaural signals after applying the methods described in Section III is computed using the signals in the time domain. This is defined as the SNR gain, which is the ratio between the SNR after applying the methods and the SNR before applying the methods. Both SNR values are

calculated at each ear. The SNR gain (G_{SNR}) is formulated as follows:

$$G_{SNR}^{r,l} = 10\log_{10}\left(\frac{\sum_{t=1}^{T}|P_{r,l}^{Md}(t)|^2 / \sum_{t=1}^{T}|P_{r,l}^{Mu}(t)|^2}{\sum_{t=1}^{T}|P_{r,l}^{d}(t)|^2 / \sum_{t=1}^{T}|P_{r,l}^{u}(t)|^2}\right), \tag{40}$$

where $P_{r,l}^{Md}(t)$ and $P_{r,l}^{Mu}(t)$ are the time-domain representations of the signals defined for each method in Eq. (39) and $P_{r,l}^{d}(t)$ and $P_{r,l}^{u}(t)$ are the representations of $P_{r,l}^{d}(k)$ and $P_{r,l}^{u}(k)$ in the time domain, respectively. Another proposed measure assesses the signal to distortion ratio (SDR) of the desired signal after applying the proposed methods, by calculating the normalized error between the true desired signal and its estimation:

$$SDR_{r,l} = 10\log_{10}\left(\frac{\sum_{t=1}^{T}|P_{r,l}^{d}(t)|^2}{\sum_{t=1}^{T}|P_{r,l}^{Md}(t) - P_{r,l}^{d}(t)|^2}\right). \tag{41}$$

VI. Objective Analysis of Performance

In this section the performance of the methods described in Section III is evaluated by using computer simulations and objective measures of performance. A microphone array recording of a speaker in a reverberant room with another noise source in the room were simulated, with the aim of studying enhancement methods that remove the noise while maintaining the spatial details in the recorded speech signal. A Monte-Carlo simulation was conducted to investigate the dependence of performance on a wide range of factors related to the acoustic scene. Only relatively close and stationary noise sources were considered in these simulations. Other noise types, such as diffuse noise and non-stationary noise sources, should be considered in future investigations, as system performance may differ from that presented here.

A. Methodology

Details of the simulation are presented in this section. The Monte-Carlo simulation was composed of 1728 realizations of the acoustic scene under different conditions. In each realization a rectangular room with reverberation time T_{60} and critical distance r_c was simulated using the image method [29]. A speaker and a noise source were positioned in this room, as detailed later. Both were represented by point sources. A spherical microphone array was positioned at (x_0, y_0, z_0), and measured both the speech and the noise signals. The microphone array, the speaker and the noise source had the same position on the z-axis ($z = 1.7$ m). The SH coefficients of the sound field ($\mathbf{a}_{nm}(t)$) around the microphone array were computed using nearly-uniform sampling with order $N = 4$ and 36 samples. The order $N = 4$ was chosen to emulate practical spherical microphone arrays such as the mh Acoustics' Eigenmike [41]. The mask $\mathbf{M}(\mathbf{\Phi}, \tau, \nu)$ from Eq. (26) was calculated for $Q = 36$ directions, and with elements of vector $\mathbf{\Phi}$ defined by the directions of the nearly-uniform samples on the sphere. It is noteworthy, that the vector $\mathbf{\Phi}$ does not need to include the DOA of the desired signal. Spatial aliasing and sensor noise were assumed to be negligible for simplicity. In the same manner, $\mathbf{a}_{nm}^{d}(t)$ and $\mathbf{a}_{nm}^{u}(t)$ were calculated separately. After representing $\mathbf{a}_{nm}^{d}(t)$ and $\mathbf{a}_{nm}^{u}(t)$ in the time-frequency domain by applying the STFT (with a Hanning window of length 512 samples, 50 % overlap, FFT of length 512 and sampling frequency of 16 kHz), the processing methods from Section III were applied by using the instantaneous estimation of the SNR (Eqs. (17), (22) and (32)). Binaural signals were then computed by using the Cologne HRTF compilation of the Neumann KU-100 [42]. The binaural signals were generated for each realization. A schematic description of the acoustic scenes is shown in Fig. 1. The independent variables in this simulation, and the values of each variable are presented in Table I. As shown in the table, the simulation was composed of four different speakers pronouncing a different utterance each, three values for SNRin, six noise types, two different rooms, three different angle sets for the speaker and the noise sources relative to the listener's head, and four distances of the sources relative to the recording array, which add up to a total of 1728 realizations for all combinations. For each realization the objective measures from Section V were calculated.

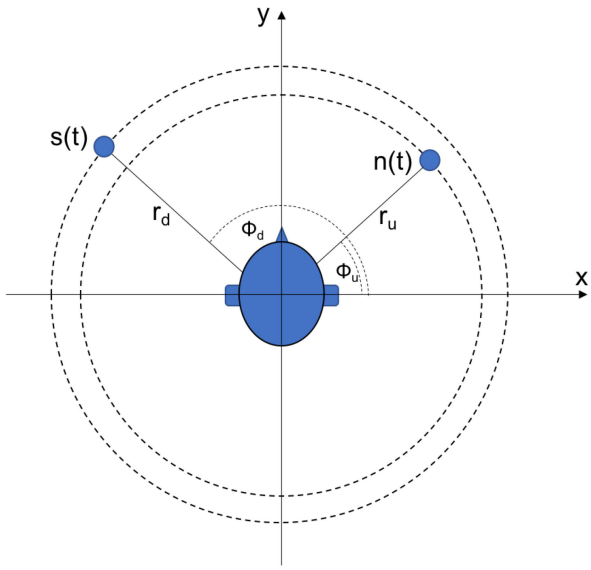


Fig. 1. Schematic of the acoustic scene, showing the two sources and the human head in the room. In the case of the recording scene, the human head is replaced by a spherical microphone array.

B. Results

The objective measures from Section V, which were calculated for each realization, were computed and are presented in this section.

1) SDR and SNR Gain: The measures G_{SNR} and SDR were computed by averaging over all conditions in the Monte-Carlo simulation, and are presented in Fig. 2. As shown in the figure, the medians of all methods for both measures differ with significance of $p < 0.05$. The TFS achieved higher median values than the TFSH for both measures, by about 2-3 dB. This implies that, in general, the TFS distorts the desired signal less and better attenuates the noise compared to the TFSH method. It is also shown in Fig. 2 that the TFSH and the TFS achieved significantly higher median values than the Beamforming method in the case of SDR, but slightly lower median values in the case of G_{SNR}. The reason for this will be discussed later.

TABLE I
DETAILS AND PARAMETERS OF THE MONTE-CARLO SIMULATION

Independent variable	# of values	Description
Speech ($s(t)$)	4	Four different utterances taken from the TIMIT corpus [43], sampled at $f_s = 16$ kHz.
SNRin	3	SNRin= $-10, 0, 10$ dB.
Noise ($n(t)$)	6	Four types of noise: faucet, air-conditioner (AC), blender and fan noise from the free sound repository Freesound.org [44] and white and pink noise generated in Matlab. All signals were sampled at $f_s = 16$ kHz.
Room	2	Two different rooms: The first room is a rectangular room of dimensions $8\,\text{m} \times 5\,\text{m} \times 3\,\text{m}$, $T_{60} = 0.7\,s$ with critical distance $r_c = 0.74\,m$. The spherical microphone array was positioned at $(x_0, y_0, z_0) = (2, 2, 1.7)$ m. The second room is a rectangular room of dimensions $4\,\text{m} \times 3\,\text{m} \times 3\,\text{m}$, $T_{60} = 0.5\,s$ with critical distance $r_c = 0.48\,m$. The spherical microphone array was positioned at $(x_0, y_0, z_0) = (1, 1.5, 1.7)$ m.
Speaker and noise source angles	3	Three different directions for the desired and the undesired sources were simulated: $(\Phi_d, \Phi_u) = (120°, 60°)$, $(\Phi_d, \Phi_u) = (150°, 90°)$ and $(\Phi_d, \Phi_u) = (90°, 30°)$, where Φ_d and Φ_u are defined in Fig. 1.
Distance	4	Four source distances (normalized by r_c), defined as the pair $(\tilde{r}_d, \tilde{r}_u)$, were simulated: $(0.5, 0.5)$, $(2, 0.5)$, $(0.5, 2)$ and $(2, 2)$, where $\tilde{r}_d = \frac{r_d}{r_c}$ and $\tilde{r}_u = \frac{r_u}{r_c}$ and r_u and r_d are defined in Fig. 1.

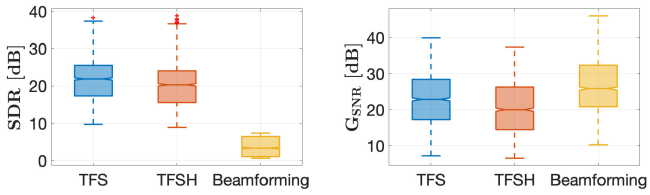


Fig. 2. The objective measures G_{SNR} (right) and SDR (left). SDR is the average of SDR_r and SDR_l, and G_{SNR} is the average of G^r_{SNR} and G^l_{SNR}. Box plot visualization: the median is the middle line; the bottom and top edges represent the 25th and 75th percentiles; the whiskers represent the extreme values, excluding outliers; the notches have been calculated such that boxes with non-overlapping notches have medians which are different at the 95% significance level. Outliers are marked with red "+".

Following the definitions of the TFS (Eq. (26)) and the TFSH (Eq. (18)) methods, it seems that a significant difference between the two methods is found in the ability of the TFS method to separate sources in the spatial domain in addition to in the time-frequency domain. Hence, it may be expected that the TFS would perform better than the TFSH when the sound field is composed of significant direct components from the sound sources. However, in the case where the sources produce highly reverberant or diffuse sound fields, the TFS method may not have that advantage over the TFSH method. In order to investigate this hypothesis, and to study the effect of source distance on performance, which may be an important factor in practical applications, performance is evaluated as a function of the independent variable "Distance". The case (0.5,0.5) in Table I, where the distance of the sources is half of the critical distance, represents a sound field with a dominant direct component,

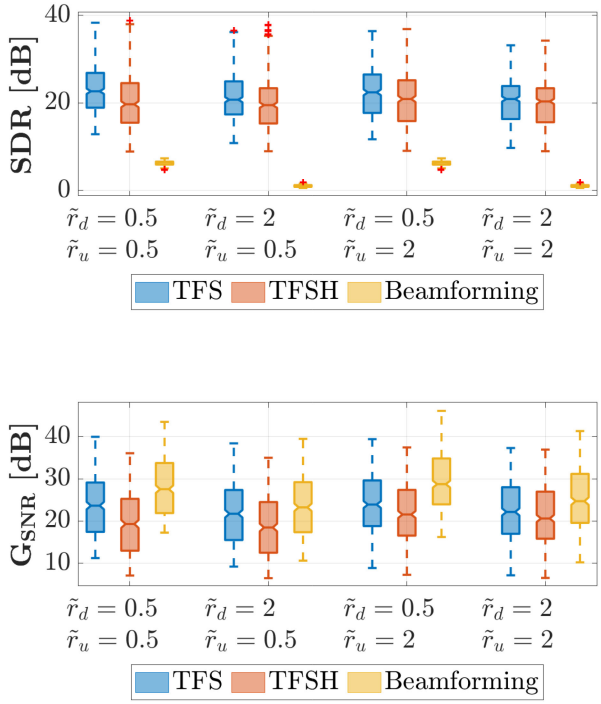


Fig. 3. The objective measures G_{SNR} (bottom) and SDR (top) as a function of the independent variable "Distance". SDR is the average of SDR_r and SDR_l, and G_{SNR} is the average of G^r_{SNR} and G^l_{SNR}.

whereas the case (2,2), where the distance of the sources is twice the critical distance, represents a sound field which is more reverberant or diffuse.

Fig. 3 presents G_{SNR} and SDR for the three methods, as a function of the independent variable "Distance," i.e. four sets of distances as presented in Table I. The results are averaged over all other factors from Table I. As shown in Fig. 3, the median of the TFS is significantly higher than the median of the TFSH ($p < 0.05$) for all the distances and for both measures, except for the "Distance" case (2,2), which corroborates the hypothesis presented above. For the "Distance" case (0.5,0.5), Fig. 3 shows that the TFS method outperforms the TFSH method by about 4-5 dB in both SNR and SDR, clearly showing its superiority for this case.

Fig. 3 also shows that the Beamforming method performs much better than the other two methods with respect to G_{SNR}, but significantly worse with respect to SDR. This can be explained by the way the Beamforming method reproduces the desired sound field - it attenuates a significant component of the noise using the beam-pattern, therefore achieving high G_{SNR}. However, the same beam pattern also attenuates important components of the desired signal, leading to poor performance with respect to SDR.

2) Residual Noise: Some applications may benefit from preserving the spatial cues of the residual noise. Examples could be traffic noise or impact noise which may require special attention from the listener. Therefore, in this section the preservation of the spatial cues of the residual noise after applying the aforementioned methods is studied, using the measures of IACC_t and the ILD [45], [46]. The IACC_t measures the correlation time of the signals at the right and the left ears and represents the inter-aural

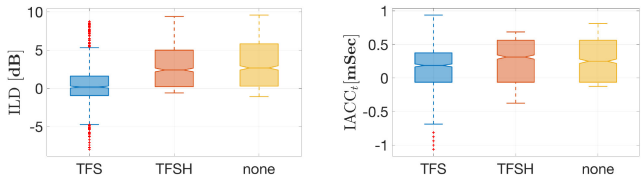


Fig. 4. The ILD (left) and IACC$_t$ (right) after applying the TFS (blue), the TFSH (red) methods and for the unprocessed noise signal (yellow).

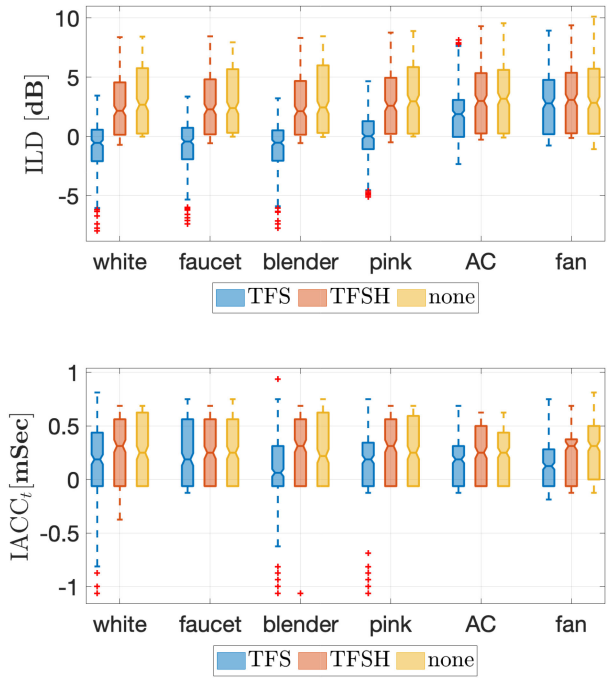


Fig. 5. ILD (top) and the IACC$_t$ (bottom) as a function of noise type.

time difference (ITD) [46], which is an important localization cue up to 1500 Hz. The ILD is an important localization cue for frequencies higher than 1500 Hz. Hence, in order to calculate the IACC$_t$, the residual noise signal in the time domain $P_{r,l}^{Mu}(t)$ (Eq. (39)) was filtered by a band-pass filter in the frequency range 100 Hz to 1500 Hz, and the ILD was calculated for frequencies above 1500 Hz and then averaged over all frequency bands.

Fig. 4 presents the results of the ILD and IACC$_t$ for the TFS and the TFSH methods compared to the unprocessed noise after averaging over all data from the Monte-Carlo simulation presented in Section VI. As shown in Fig. 4, for both cues the median of the TFSH and the unprocessed noise are not significantly different, whereas the median of the TFS is significantly different from the TFSH and from the unprocessed noise signal with significance of $p < 0.05$. This result implies that the TFSH method may better preserve the spatial cues of the residual noise. Fig. 4 shows the averaged performance under all experimental conditions, including different noise types. It may be interesting to investigate whether the level to which a method maintains spatial cues of the noise depends on the noise type. Therefore, ILD and IACC$_t$ were evaluated with respect to the independent variable "Noise" from Table I, i.e. six type of noise sources. Fig. 5 presents the results. As presented in Fig. 5, in the case of both ILD and IACC$_t$ the medians of the TFSH and the medians of the unprocessed noise are not significantly different for all noise types. This may suggest that the TFSH may preserve the DOA of the noise source independently of the noise type. On the other hand, the medians of the TFS method differ from the medians of the unprocessed noise with significance of $p < 0.05$ for all noise types, except for the faucet noise (with respect to IACC$_t$), and the fan noise (with respect to ILD). This may imply that the TFS may not, in general, preserve the DOA of the noise source. However, the preservation of spatial information may change between noise types, and so the actual perception of direction may differ with noise type. To evaluate the latter, a listening test is performed and reported in Section VIII.

VII. Listening Test 1 - Enhanced Desired Signal

In this section, we report on a listening test that was conducted to further study the ability of the processing methods from Section III to preserve the spatial information of the desired source, while attenuating the contribution of the undesired source to the sound field.

A. Methodology

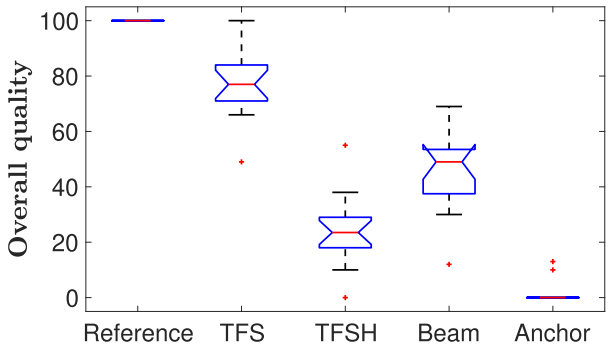
Binaural signals were generated in two acoustic scenes that were sampled from the Monte-Carlo simulations. A schematic description of the acoustic scene is shown in Fig. 1. The parameters presented in Fig. 1 and their values for each of the acoustic scenes are described in Table I. Both acoustic scenes include the following parameters: $s(t)$ - a single female speaker, $\text{SNR}_{in} = 0\,\text{dB}$, $n(t)$ - pink noise, Room - the first room in Table I, $(\Phi_d, \Phi_u) = (120°, 60°)$ and $(\tilde{r}_d, \tilde{r}_u) = (0.5, 0.5)$. The difference between the acoustic scenes is the distance between the sources and the listener's head. In the first acoustic scene, the distance between the sources and the listener's head is half of the critical distance $(\tilde{r}_d, \tilde{r}_u) = (0.5, 0.5)$ and in the second acoustic scene this distance is twice the critical distance $(\tilde{r}_d, \tilde{r}_u) = (2, 2)$.

B. Experimental Setup

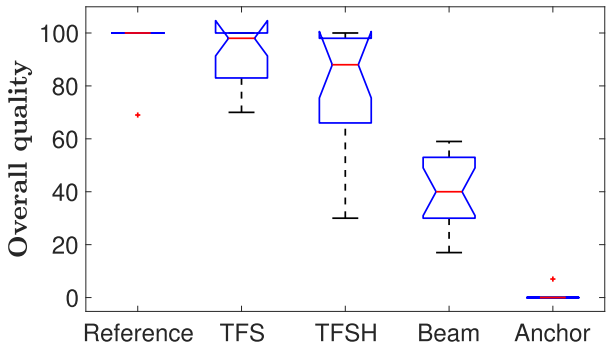
A listening test based on Recommendation ITU-R BS.1534-1 (MUSHRA, MUltiple Stimuli with Hidden Reference and Anchor) [47] was conducted. For both acoustic scenes, five binaural signals were generated:[1]

1) **Reference**: a binaural signal generated only by the desired signal ($P_{r,l}^{d}(t)$).
2) **TFS**: a binaural signal generated after applying the TFS method ($P_{r,l}^{M}(t)$).
3) **TFSH**: a binaural signal generated after applying the TFSH method ($P_{r,l}^{M}(t)$).
4) **Beamforming**: a binaural signal generated after applying the beamforming method ($P_{r,l}^{M}(t)$).
5) **Anchor**: the desired source signal plus the undesired source signal as they are measured at the center of the microphone array, without any processing. These signals can be calculated by using the zero SH coefficients of the

[1]This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors. This includes the signals of the two listening tests and a readme file. This material is 6.8 MB in size.



(a) Acoustic scene 1.



(b) Acoustic scene 2.

Fig. 6. Results for the overall quality ratings in scene 1 (a) and scene 2 (b).

desired and undesired signals, namely $a_{00}^{d}(t)$ and $a_{00}^{u}(t)$, respectively.

All signals were played back using the Matlab (MATLAB R2018b) audio recorder and AKG K702 headphones. 16 normal hearing subjects participated in this experiment. The experiment included 2 MUSHRA screens - one screen for each acoustic scene as detailed above. For both MUSHRA screens, each participant was asked to rate the overall quality of the signals relative to the reference signal, on a scale from 0 to 100. The overall quality was defined as a combination of the following properties: Externalization, Localization, Envelopment, Noise-like artifact and Distortion, as defined in [48]. Before rating, the participants performed a training task in order to ensure that the instructions were clearly understood and to familiarize the participants with the stimuli.

C. Results

The overall quality results are presented in Fig. 6 for both acoustic scenes.

As shown in Fig. 6, in acoustic scene 1 the median scores of all signals differ with significance $p < 0.05$. In this acoustic scene, the median of the TFSH method is 23.5, which places this method in the lowest place compared to the TFS method (with median 77) and the Beamforming method (with median 49). In acoustic scene 2 the median scores of all signals differ with significance $p < 0.05$, except for the TFS and the Reference scores, and the TFS and the TFSH scores. The TFS and the TFSH methods are highly rated with medians of 98 and 88, respectively. These results of the TFS and the TFSH are with correlation to the objective results in Section VI (Fig. 3), as the TFSH and the TFS methods perform approximately the same in distance case (2,2) and the TFS method performs much better than the TFSH method in distance case (0.5,0.5) for both measures (G_{SNR} and SDR).

A two-way analysis of variance (ANOVA) was used to examine the effect of the processing methods, the acoustic scenes and the interaction between them. The results indicated a significant effect for the processing methods ($p < 0.01$), for the acoustic scenes ($p < 0.01$) and for the interaction ($p < 0.05$), suggesting that the participants rated the processing methods differently. Moreover, the significant interaction between the processing methods and the acoustic scenes can be specifically explained by the TFSH method, which is rated high in acoustic scene 2 and rated low in acoustic scene 1. This change significantly impacts the interaction, showing the sensitivity of this method to the source's distance.

VIII. Listening Test 2 - Residual Noise

In this section, we report on a listening test that was conducted to further study the ability of the processing methods from Sec. III to preserve the DOA of the residual noise.

A. Methodology

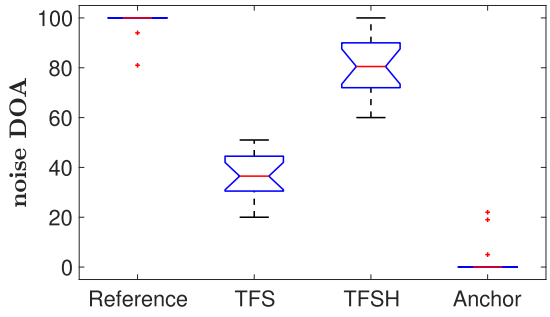
Binaural signals were generated in two acoustic scenes that were sampled from the Monte-Carlo simulations presented in Sec. VI. A schematic description of the acoustic scene is shown in Fig. 1. The parameters presented in Fig. 1 and their values for each of the acoustic scenes are described in Table I. Both acoustic scenes include the following parameters: $s(t)$ - a single female speaker, $\text{SNR}_{in} = 0\,\text{dB}$, Room - the first room in Table I, $(\Phi_d, \Phi_u) = (120^\circ, 60^\circ)$ and $(\tilde{r}_d, \tilde{r}_u) = (0.5, 0.5)$. The difference between the acoustic scenes is the noise type. In the first acoustic scene, the noise, $n(t)$, is white noise and in the second acoustic scene the noise type is fan noise.

B. Experimental Setup

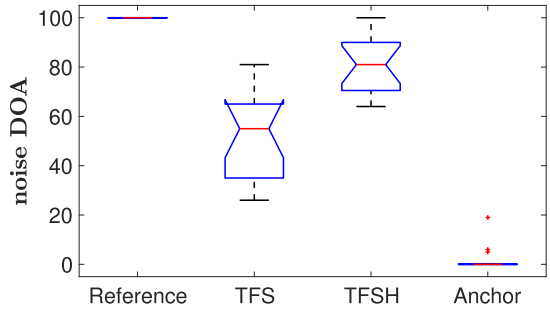
For both acoustic scenes, four binaural signals were generated:

1) **Reference**: a binaural signal generated by the unprocessed noise signal ($P_{r,l}^{u}(t)$).
2) **TFS**: a binaural signal of the residual noise generated after applying the TFS method ($P_{r,l}^{Mu}(t)$).
3) **TFSH**: a binaural signal of the residual noise generated after applying the TFSH method ($P_{r,l}^{Mu}(t)$).
4) **Anchor**: a binaural signal of the noise source, but when relocated to the position of the desired source.

All signals were played back using the Matlab (MATLAB R2018b) audio recorder and AKG K702 headphones. 16 normal hearing subjects participated in this experiment. The experiment, based on the MUSHRA test, included 2 screens - one screen for each acoustic scene as detailed above. For both MUSHRA screens, each participant was asked to rate the similarity to the reference signal, based on the DOA, on a scale from 0 to 100. Before rating, the participants performed a training task, in order



(a) White noise (scene 1).



(b) Fan noise (scene 2).

Fig. 7. Results for the residual noise's DOA ratings in scene 1 (a) and scene 2 (b).

to ensure that the instructions were clearly understood and to familiarize the participants with the stimuli.

C. Results

The results are presented in Fig. 7 for both acoustic scenes. As shown in Fig. 7, in acoustic scenes 1 and 2 the median scores of all signals differ with significance $p < 0.05$. In acoustic scene 1, the median of the TFS method is 36.5, which is significantly lower than the TFSH method (with median 80.5). In acoustic scene 2, the median of the TFS method is 55, and is also significantly lower compared to the TFSH method (with median 81). These results of the TFS and the TFSH methods are in agreement with the objective results in Section VI (Fig. 5), as the TFSH method seems to better preserve the DOA of the residual noise compared to the TFS method for both noise types, although actual performance may depend on noise type.

A two-way analysis of variance (ANOVA) was used to examine the effect of the processing methods, the acoustic scenes and the interaction between them. The results indicated a significant effect for the processing methods ($p < 0.01$), for the acoustic scenes ($p < 0.01$) and for the interaction ($p < 0.01$), suggesting that the participants rated the processing methods differently. Moreover, the significant interaction between the processing methods and the acoustic scenes can be specifically explained by the TFS method, which is rated higher in acoustic scene 2 than in acoustic scene 1. This change significantly impacts the interaction, showing the dependence on the residual noise type of the TFS method in preserving the DOA. It is noteworthy that the participants in the listening test rated the DOA of the residual noise by listening to the residual noise only. However, in the actual noise reduced signal, the residual noise may be partially masked by the desired sound, which may affect how it is perceived.

IX. Conclusion

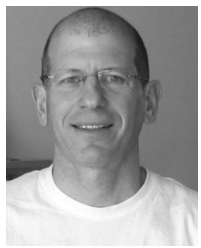
According to the subjective and the objective analyses presented in this paper, the TFS method seems to outperform the TFSH method in terms of preserving the spatial information of the desired signal. It was specifically shown that there is a strong dependence of the performance of the TFSH method on the sources' distances from the microphone array, which is a drawback of this method. Nevertheless, for highly reverberant environments with distant sources this method performs approximately the same as the TFS mask in terms of SDR and SNR gain, and may be preferable due to the direct processing in the SH domain. Moreover, it was shown that the TFSH method preserves the DOA of the residual noise better than the TFS method. It was specifically shown that the DOA of the residual noise after applying the TFS method is dependent on the noise type. The Beamforming method was found to be very successful in noise reduction but very poor in preserving the spatial information of the desired sound. However, as this study examined only a relatively small set of audio signals and acoustic conditions, a more comprehensive study should be performed to generalize these conclusions.

References

[1] T. S. Perry, "Virtual reality goes social," *IEEE Spectrum*, vol. 53, no. 1, pp. 56–57, Jan. 2016.
[2] C. Moore, "The Virtual Yellow House: Experimental tangling with virtual reality," *IEEE Consum. Electron. Mag.*, vol. 5, no. 4, pp. 103–104, Oct. 2016.
[3] B. Markwalter, "Entertainment and immersive content: What's in store for your viewing pleasure," *IEEE Consum. Electron. Mag.*, vol. 4, no. 1, pp. 83–86, Jan. 2015.
[4] S. Greenwald *et al.*, "Technology and applications for collaborative learning in virtual reality," presented at the Making a Difference: Prioritizing Equity and Access in CSCL, 12th Int. Conf. Comput. Supported Collaborative Learn., 2017.
[5] S. Mehrotra, W.-G. Chen, Z. Zhang, and P. A. Chou, "Realistic audio in immersive video conferencing," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2011, pp. 1–4.
[6] V. A. Nguyen, J. Lu, S. Zhao, D. L. Jones, and M. N. Do, "Teleimmersive audio-visual communication using commodity hardware [applications corner]," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 118–136, Nov. 2014.
[7] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook Array Processing and Sensor Networks*. New York, NY, USA: Wiley, 2010, pp. 269–302.
[8] 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 Process. Mag.*, vol. 32, no. 2, pp. 18–30, Mar. 2015.
[9] E. Hadad, S. Doclo, and S. Gannot, "The binaural LCMV beamformer and its performance analysis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 543–558, Mar. 2016.
[10] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating interactive virtual acoustic environments," *J. Audio Eng. Soc.*, vol. 47, no. 9, pp. 675–705, 1999.
[11] A. Avni, J. Ahrens, M. Geier, S. Spors, H. Wierstorf, and B. Rafaely, "Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution," *J. Acoust. Soc. Amer.*, vol. 133, no. 5, pp. 2711–2721, 2013.
[12] Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, and B. Rafaely, "Spectral equalization in binaural signals represented by order-truncated spherical harmonics," *J. Acoust. Soc. Amer.*, vol. 141, no. 6, pp. 4087–4096, 2017.

[13] T. D. Abhayapala and D. B. Ward, "Theory and design of high order sound field microphones using spherical microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, vol. 2, pp. 1949–1952.
[14] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 135–143, Jan. 2005.
[15] D. P. Jarrett, E. A. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing*. Berlin, Germany: Springer-Verlag, 2017, pp. 11–22.
[16] E. Fernandez-Grande and A. Xenaki, "Compressive sensing with a spherical microphone array," *J. Acoust. Soc. Amer.*, vol. 139, no. 2, pp. EL45–EL49, 2016.
[17] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 1562–1566.
[18] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2136–2147, Dec. 2015.
[19] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. Hoboken, NJ, USA: Wiley, 2004.
[20] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, vol. 1. Berlin, Germany: Springer-Verlag, 2008.
[21] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2449–2464, Dec. 2015.
[22] A. H. Moore, L. Lightburn, W. Xue, P. A. Naylor, and M. Brookes, "Binaural mask-informed speech enhancement for hearing aids with head tracking," in *Proc. 16th Int. Workshop Acoust. Signal Enhancement*, 2018, pp. 461–465.
[23] N. R. Shabtai and B. Rafaely, "Generalized spherical array beamforming for binaural speech reproduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 1, pp. 238–247, Jan. 2014.
[24] M. Jeffet, N. R. Shabtai, and B. Rafaely, "Theory and perceptual evaluation of the binaural reproduction and beamforming tradeoff in the generalized spherical array beamformer," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 4, pp. 708–718, Apr. 2016.
[25] C. Borrelli, A. Canclini, F. Antonacci, A. Sarti, and S. Tubaro, "A denoising methodology for higher order ambisonics recordings," in *Proc. 16th Int. Workshop Acoust. Signal Enhancement*, 2018, pp. 451–455.
[26] U. Abend and B. Rafaely, "Spatio-spectral masking for spherical array beamforming," in *Proc. IEEE Int. Conf. Sci. Elect. Eng.*, 2016, pp. 1–5.
[27] A. Herzog and E. A. Habets, "Direction preserving wiener matrix filtering for ambisonic input-output systems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 446–450.
[28] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 260–276, Feb. 2010.
[29] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.
[30] B. Rafaely, *Fundamentals of Spherical Array Processing*, vol. 8. Berlin, Germany: Springer, 2015.
[31] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1494–1505, Oct. 2014.
[32] D. Khaykin and B. Rafaely, "Coherent signals direction-of-arrival estimation using a spherical microphone array: Frequency smoothing approach," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2009, pp. 221–224.
[33] B. Rafaely, "Plane-wave decomposition of the sound field on a sphere by spherical convolution," *J. Acoust. Soc. Amer.*, vol. 116, no. 4, pp. 2149–2157, 2004.
[34] Y. A. Huang and J. Benesty, *Audio Signal Processing for Next-Generation Multimedia Communication Systems*. Berlin, Germany: Springer, 2007.
[35] N. Madhu, A. Spriet, S. Jansen, R. Koning, and J. Wouters, "The potential for speech intelligibility improvement using the ideal binary mask and the ideal Wiener filter in single channel noise reduction systems: Application to auditory prostheses," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 1, pp. 63–72, Jan. 2013.
[36] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication," *Speech Commun.*, vol. 53, no. 5, pp. 677–689, 2011.
[37] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Conf. Proc.*, 1996, vol. 2, pp. 629–632.
[38] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.
[39] J. Sheaffer, M. Van Walstijn, B. Rafaely, and K. Kowalczyk, "Binaural reproduction of finite difference simulations using spherical array processing," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2125–2135, Dec. 2015.
[40] B. Rafaely and A. Avni, "Interaural cross correlation in a sound field represented by spherical harmonics," *J. Acoust. Soc. Amer.*, vol. 127, no. 2, pp. 823–828, 2010.
[41] M. Acoustics, "Em32 eigenmike microphone array release notes (v17. 0)," Summit, NJ, USA, 2013.
[42] B. Bernschütz, "A spherical far field hrir/hrtf compilation of the neumann ku 100," in *Proc. 40th Italian (AIA) Annu. Conf. Acoust. 39th German Annu. Conf. Acoust. Conf. Acoust.*, 2013, pp. 592–595.
[43] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Darpa TIMIT acoustic-phonetic continuous speech corpus cd-rom TIMIT," Nat. Inst. Standards Technol. (NIST), Tech. Rep. NISTIR 4930, 1993.
[44] "Freesound," Apr. 2013. [Online]. Available: http://www.freesound.org
[45] "Acoustics: Measurement of the reverberation time of rooms with reference to other acoustical parameters," International Organization for Standardization, Geneva, Switzerland, 1997.
[46] J. Blauert, *The Technology of Binaural Listening*. Berlin, Germany: Springer-Verlag, 2013.
[47] "1534-1, 'method for the subjective assessment of intermediate quality levels of coding systems (mushra),"' International Telecommunication Union, Geneva, Switzerland, 2003.
[48] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl, "A spatial audio quality inventory (SAQI)," *Acta Acustica United Acustica*, vol. 100, no. 5, pp. 984–994, 2014.

Moti Lugasi (Student Member, IEEE) received the B.Sc. (*cum laude*) and M.Sc. degrees in electrical and computer engineering in 2017 and 2020, respectively, from the Ben-Gurion University of the Negev, Beer-Sheva, Israel, where he is currently working toward the Ph.D. degree. His current research interests focus on speech processing with microphone arrays under real-life conditions. He is the recipient of the Ben-Gurion University Darom-Lachish fellowship.



Boaz Rafaely (Senior Member, IEEE) received the B.Sc. degree (*cum laude*) in electrical engineering from Ben-Gurion University, Beer-Sheva, Israel, in 1986, the M.Sc. degree in biomedical engineering from Tel-Aviv University, Tel Aviv, Israel, in 1994, and the Ph.D. degree from the Institute of Sound and Vibration Research (ISVR), Southampton University, Southampton, U.K., in 1997. At the ISVR, he was appointed as a Lecturer in 1997 and a Senior Lecturer in 2001, working on active control of sound and acoustic signal processing. In 2002, he spent six months as a Visiting Scientist with the Sensory Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology (MIT), Cambridge, investigating speech enhancement for hearing aids. He then joined the Department of Electrical and Computer Engineering, Ben-Gurion University as a Senior Lecturer in 2003, and appointed as an Associate Professor in 2010, and a Professor in 2013. He is currently heading the acoustics laboratory, investigating methods for audio signal processing and spatial audio. During 2010–2014, he has served as an Associate Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, and during 2013–2018 as a member of the IEEE Audio and Acoustic Signal Processing Technical Committee. He also served as an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS during 2015–2019, *IET Signal Processing* during 2016–2019, and currently for *Acta Acustica*. During 2013–2016, he has served as the Chair of the Israeli Acoustical Association, and is currently chairing the Technical Committee on Audio Signal Processing in the European Acoustical Association. He was awarded the British Council's Clore Foundation Scholarship.