Audio Engineering Society

Conference Paper

Presented at the AES International Conference on
Audio for Virtual and Augmented Reality
2020 August 17 – 19, Online

Enhancement of Ambisonics signals using time-frequency masking

Moti Lugasi and Boaz Rafaely

School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel

Correspondence should be addressed to Moti Lugasi (motilu@post.bgu.ac.il)

ABSTRACT

Spatial audio is an essential part of virtual reality. Unlike synthesized signals, spatial audio captured in the real world may suffer from background noise which degrades the quality of the signals. While some previous works have addressed this problem, and suggested methods to attenuate the undesired signals while preserving the desired signals with minimum distortion, these only succeed partially. Recently, methods aiming to achieve preservation of the desired signal in its entirety have been proposed, and in this work we study such methods that are based on time-frequency masking. Two masks were investigated: one in the spherical harmonics (SH) domain, and the other in the plane wave density (PWD) function domain, referred to here as the spatial domain. These two methods were compared with a low-end reference method that uses a single maximum directivity beamformer followed by a single channel time-frequency mask. A subjective investigation was conducted to estimate the performance of these methods, and showed that the spatial mask preserves the desired sound field better, while the SH mask preserves the spatial cues of the residual noise better.

1 Introduction

Spatial audio that incorporates spatial information in the reproduction and playback of sound has become increasingly popular in recent years in a variety of applications. One such application is virtual reality, featuring in distance learning, gaming and entertainment, architectural design, and more [1, 2, 3, 4]. Although spatial audio signals can be generated artificially, capturing these signals in the real world by microphone arrays is of great importance in applications such as the recording of music events, communication in video conferencing meetings, and for hearing aids [5, 6, 7]. Captured audio in the real world may include, in addition to desired components such as speech or music, undesired components such as noise or other interferences. It may therefore be helpful to attenuate the undesired components without distorting the spatial information in the desired components. Despite the severity of this problem, only a few studies have been published in the literature that propose solutions to this challenge.

Approaches for the attenuation of noise in hearing aid applications are based on the multi-channel Wiener fil-

ter (MWF) [8, 9, 10, 11, 12] and time-frequency masking [13, 14]. However, while useful for hearing aids, the main drawback of these methods is that the incorporation of head tracking is not possible. Hence, these methods are not suitable for virtual reality applications.

Ambisonics signals, on the other hand, can be easily used to reproduce the binaural signals and also incorporate head tracking [15]. The authors in [16, 17] suggest a method to attenuate the sum of signals per DOA of the sound field without changing the acoustic scene, so that directional interferences can be attenuated effectively. However, this method fails when the undesired signal is an ambient noise. Another method for enhancing Ambisonics signals proposes estimation of the source signal by using a maximum directivity beamformer, and using this estimation to estimate the desired sound field [18]. However, because this method assumes a dominant source, performance may degrade when the source is far from the array. In [19] the authors proposed reproduction of a directional sound field by estimating the sources with a MWF. This method may fails as well when the sound field is diffuse.

In this paper, methods that aim to preserve the entire desired sound field, using time-frequency masking, were investigated. These include: (i) the spherical harmonics (SH) mask approach from [20], which is extended to the time-frequency domain; (ii) the second approach outlined in [19], which is based on a spatial time-frequency mask; and (iii) a third method, chosen as a low-end reference and motivated by [21], based on a single beamformer with a time-frequency mask. Using listening tests, it was shown that the method from (ii) preserves the desired sound better than the method from (i), while the method from (i) preserves the DOA of the residual noise better than the method in (ii). Parts of this paper have been recently included in an extended and more detailed description of this research [22].

2 Signal model

In this section the signal model in the SH domain (Ambisonics signals) and in the plane wave density (PWD) domain (spatial domain) are presented in the short time Fourier transform (STFT) domain.

Assume the signal model of a captured sound-field represented in the Ambisonics domain is:

$$\mathbf{a}_{nm}(\tau,\nu) = \mathbf{a}^d_{nm}(\tau,\nu) + \mathbf{a}^u_{nm}(\tau,\nu), \tag{1}$$

where $\mathbf{a}^d_{nm}(\tau,\nu)$ and $\mathbf{a}^u_{nm}(\tau,\nu)$ are $(N+1)^2 \times 1$ vectors that represent the desired and undesired Ambisonics signals in the time-frequency domain, where τ is the time frame and ν is the frequency index. Using the SH matrix, which is defined as:

$$\mathbf{Y}(\overline{\Phi}) = \begin{bmatrix} \mathbf{y}^T(\Phi_1) \\ \mathbf{y}^T(\Phi_2) \\ \vdots \\ \mathbf{y}^T(\Phi_Q) \end{bmatrix}, \tag{2}$$

where $\mathbf{y}(\Phi_l) = [Y_0^0(\Phi_l), Y_1^{-1}(\Phi_l) Y_1^0(\Phi_l), ..., Y_N^N(\Phi_l)]^T$, $\overline{\Phi} = [\Phi_1, \Phi_2, ..., \Phi_Q]$ is a vector of Q arbitrary directions and $Y_n^m(\Phi)$ denotes the SH functions of order n and degree m [15], the signal model in the PWD domain is given by:

$$\mathbf{a}(\overline{\Phi},\tau,\nu) = \mathbf{Y}(\overline{\Phi})\mathbf{a}_{nm}(\tau,\nu) = \mathbf{a}^d(\overline{\Phi},\tau,\nu) + \mathbf{a}^u(\overline{\Phi},\tau,\nu), \tag{3}$$

where $\mathbf{a}^d(\overline{\Phi},\tau,\nu)$ and $\mathbf{a}^u(\overline{\Phi},\tau,\nu)$ are $Q \times 1$ vectors of the desired and the undesired signals in the PWD domain.

3 Wiener masking methods

In this section, three masking methods for noise reduction are presented.

3.1 Time-Frequency-Spherical Harmonics mask (TFSH mask)

The TFSH mask is applied to the Ambisonics signals in the STFT domain. The Wiener mask is defined as:

$$M(n,m,\tau,\nu) = \frac{SNR(n,m,\tau,\nu)}{SNR(n,m,\tau,\nu)+1}, \tag{4}$$

while here as well the instantaneous SNR is calculated using oracle information:

$$SNR(n,m,\tau,\nu) = \frac{|a^d_{nm}(\tau,\nu)|^2}{|a^u_{nm}(\tau,\nu)|^2}, \tag{5}$$

where $a^d_{nm}(\tau,\nu)$ and $a^u_{nm}(\tau,\nu)$ are the nm'th elements of the vectors $\hat{\mathbf{a}}^d_{nm}(\tau,\nu)$ and $\hat{\mathbf{a}}^u_{nm}(\tau,\nu)$ from Eq. (1), respectively. 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:

$$\hat{\mathbf{a}}^d_{nm}(\tau,\nu) = \mathbf{M}(n,m,\tau,\nu)\mathbf{a}_{nm}(\tau,\nu), \tag{6}$$

where $\mathbf{M}(n,m,\tau,\nu)$ is a $(N+1)^2 \times (N+1)^2$ diagonal matrix defined by using Eq. (4) as:

$$\mathbf{M}(n,m,\tau,\nu) = \text{diag}(M(0,0,\tau,\nu), M(1,(-1),\tau,\nu), ..., M(N,N,\tau,\nu)). \tag{7}$$

3.2 Time-Frequency-Space mask (TFS mask)

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

$$M(\Phi_q,\tau,\nu)=\frac{SNR(\Phi_q,\tau,\nu)}{SNR(\Phi_q,\tau,\nu)+1}, \tag{8}$$

while here the instantaneous SNR is calculated using oracle information:

$$SNR(\Phi_q,\tau,\nu)=\frac{|a_d(\Phi_q,\tau,\nu)|^2}{|a_u(\Phi_q,\tau,\nu)|^2}, \tag{9}$$

where $a_d(\Phi_q,\tau,\nu)$ and $a_u(\Phi_q,\tau,\nu)$ are the q'th elements of the vectors $\mathbf{a}^d(\overline{\Phi},\tau,\nu)$ and $\mathbf{a}^u(\overline{\Phi},\tau,\nu)$ from Eq. (3), respectively. A diagonal $Q\times Q$ matrix of the Wiener mask in Eq. (8), which is calculated for $q=1,...,Q$, can be defined as:

$$\widetilde{\mathbf{M}}(\overline{\Phi},\tau,\nu)= \\ \mathrm{diag}(M(\Phi_1,\tau,\nu),M(\Phi_2,\tau,\nu),...,M(\Phi_Q,\tau,\nu)). \tag{10}$$

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

$$\hat{\mathbf{a}}_d(\overline{\Phi},\tau,\nu)=\widetilde{\mathbf{M}}(\overline{\Phi},\tau,\nu)\mathbf{a}(\overline{\Phi},\tau,\nu). \tag{11}$$

By using Eq. (3), Eq. (11) can be represented in the SH domain as:

$$\hat{\mathbf{a}}^d_{nm}(\tau,\nu)=\mathbf{M}(\overline{\Phi},\tau,\nu)\mathbf{a}_{nm}(\tau,\nu). \tag{12}$$

where $\mathbf{M}(\overline{\Phi},\tau,\nu)=\mathbf{Y}^{\dagger}(\overline{\Phi})\widetilde{\mathbf{M}}(\overline{\Phi},\tau,\nu)\mathbf{Y}(\overline{\Phi})$ and $\mathbf{Y}^{\dagger}(\overline{\Phi})=[\mathbf{Y}^H(\overline{\Phi})\mathbf{Y}(\overline{\Phi})]^{-1}\mathbf{Y}^H(\overline{\Phi})$.

3.3 Beamforming followed by masking

As a low-end reference to the other methods, a method that uses beamforming and a time-frequency mask is suggested. In this way, only the spatial information of the direct sound is preserved.

By applying a maximum directivity beamformer in the DOA of the desired source (Φ_s) in the SH domain, the array output is given by:

$$z(\tau,\nu)=\mathbf{y}^T(\Phi_s)\mathbf{a}_{nm}(\tau,\nu)=z_d(\tau,\nu)+z_u(\tau,\nu), \tag{13}$$

where $\mathbf{y}(\Phi_s)$ is defined in Eq. (2), and $z_d(\tau,\nu)=\mathbf{y}^T(\Phi_s)\mathbf{a}^d_{nm}(\tau,\nu)$ and $z_u(\tau,\nu)=\mathbf{y}^T(\Phi_s)\mathbf{a}^u_{nm}(\tau,\nu)$ are the desired and the undesired signals at the output of the beamformer, respectively. In order attenuate the residual noise in the source DOA (Φ_s), a Wiener mask is applied to $z(\tau,\nu)$ in the time-frequency domain. The Wiener mask is defined as:

$$M(\tau,\nu)=\frac{SNR(\tau,\nu)}{SNR(\tau,\nu)+1}, \tag{14}$$

while here as well the instantaneous SNR is calculated using oracle information:

$$SNR(\tau,\nu)=\frac{|z_d(\tau,\nu)|^2}{|z_u(\tau,\nu)|^2}. \tag{15}$$

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

$$\hat{z}_d(\tau,\nu)=M(\tau,\nu)z(\tau,\nu). \tag{16}$$

4 Binaural reproduction

In order to reproduce the binaural signal due to the desired sound field, the sound pressure at the right ear, $P_r(k)$, and the left ear, $P_r(k)$, at wave-number k, can be calculated approximately using the following equation [23]:

$$P_{r,l}(k)\cong\sum_{n=0}^{N}\sum_{m=-n}^{n}[\tilde{a}_{nm}(k)]^*H^{r,l}_{nm}(k), \tag{17}$$

where $\tilde{a}_{nm}(k)=(-1)^m[a_{n(-m)}(k)]^*$ and $H^{r,l}_{nm}$ are the SH coefficients of the head related transfer function (HRTF) of the left and the right ears. For TFS and TFSH masks, the binaural signal, denoted as $P^M_{r,l}(k)$, is reproduced by using an estimator of the desired signal represented in the frequency domain ($\hat{\mathbf{a}}^d_{nm}(k)$) and Eq. (17).

For the beamforming method, due to the single channel output, as in Eq. (13), only the HRTF in the desired source DOA (Φ_s) is used:

$$P^M_{r,l}(k)=\hat{z}_d(k)H_{r,l}(\Phi_s,k), \tag{18}$$

where $\hat{z}_d(k)$ is the representation of Eq. (16) in the frequency domain.

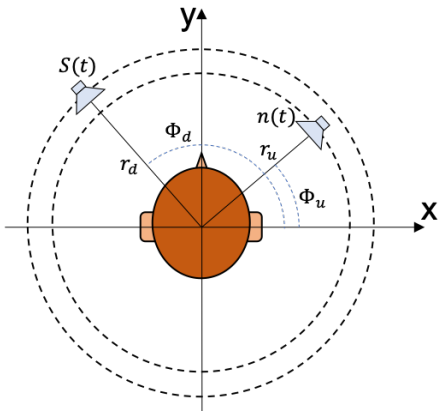


Fig. 1: Schematic of the acoustic scene, showing the two sources and the human head in the room.

5 Listening test 1 - enhanced desired signal

Two listening tests were conducted. The first focused on the capacity of the proposed methods 3 to preserve the spatial information of the desired signal, while attenuating the undesired signals, and is described below.

5.1 Setup

Two different acoustic scenes, shown schematically in Fig. 1, were the setting for generating the binaural signals. The parameters were: $s(t)$ - a desired source signal due to a single female speaker, $\mathrm{SNR}_{in} = 0\,\mathrm{dB}$, $n(t)$ - an undesired pink noise, room - a rectangular room of dimensions $8\,\mathrm{m} \times 5\,\mathrm{m} \times 3\,\mathrm{m}$, with reverberation time $T_{60} = 0.7\,s$ and critical distance $r_c = 0.74\,m$, $(\Phi_d, \Phi_u) = (120^\circ, 60^\circ)$ and $(r_d, r_u) = (0.5r_c, 0.5r_c)$. The difference between the two scenes lies in the distance between the sources and the listener's head, where in the first scene this distance is half of the critical distance $(r_d, r_u) = (0.5r_c, 0.5r_c)$, and in the second this distance twice the critical distance $(r_d, r_u) = (2r_c, 2r_c)$.

5.2 Methodology

A listening test was conducted for both scenes, using the protocol found in Recommendation ITU-R BS.1534-1 (MUSHRA, MUltiple Stimuli with Hidden Reference and Anchor) [24]. The following five binaural signals were generated:

1. **Reference**: a binaural signal generated only by the desired signal.
2. **TFS**: a binaural signal generated following the application of the TFS method.
3. **TFSH**: a binaural signal generated following the application of the TFSH method.
4. **Beamforming**: a binaural signal generated following the application of the beamforming method.
5. **Anchor**: a sum of the desired and the undesired signals, unprocessed, as they are measured at the center of the microphone array.

All signals were played back using the Matlab (MATLAB R2018b) audio recorder and AKG K702 headphones. 16 normal hearing subjects participated in this experiment. For each MUSHRA screens, the participants were asked to rate the overall quality of the signals relative to the reference signal. The definition of overall quality comprised five features: Externalization, Localization, Envelopment, Noise-like artifact and Distortion, as defined in [25].

5.3 Results

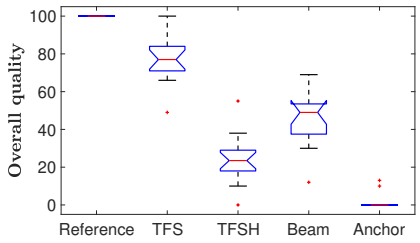
Fig. 2 presents the overall quality results for the two different scenes. It can be seen that the median scores of all signals differ with significance $p < 0.05$ in scenes 1. The median of the TFSH method is much lower that for the TFS method (23.5 compared to 77), and also worse than for the Beamforming method (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.

6 Listening test 2 - residual noise

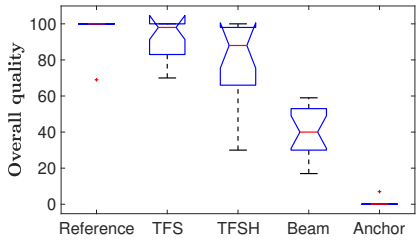
The second of the two listening tests focused on the capacity of the proposed methods (Sec. 3) to preserve the DOA of the residual noise, and is described below.

6.1 Setup

The experimental setup and parameters for the second test were the same as for the first, described in the preceding section, except for the noise signal, which was changed for each scene: white noise and fan noise, respectively, for the first and second scenes. For both scenes $(r_d, r_u) = (0.5r_c, 0.5r_c)$.

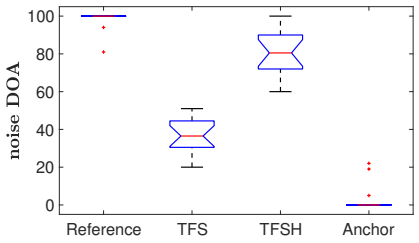


(a) Acoustic scene 1.

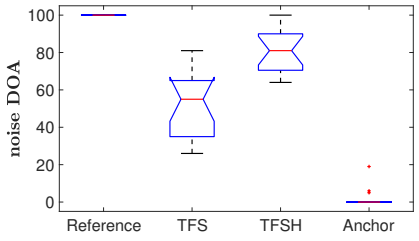


(b) Acoustic scene 2.

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



(a) White noise (scene 1).



(b) Fan noise (scene 2).

Fig. 3: Results for the residual noise's DOA ratings in scene 1 (a) and scene 2 (b).

6.2 Methodology

In the second experiment the same setup was used, but four (not five) binaural signals were generated for each acoustic scene, as follows:

1. **Reference**: a binaural signal generated by the unprocessed noise signal.
2. **TFS**: a binaural signal of the residual noise generated following the application of the TFS method.
3. **TFSH**: a binaural signal of the residual noise generated following the application of the TFSH method.
4. **Anchor**: a binaural signal of the noise source, but relocated to the position of the desired source.

In this listening test, two screens of the MUSHRA test were generated, where each screen was applied for one of the two acoustic scenes described in the preceding section.

6.3 Results

Fig. 3 presents the results for the two different scenes. It can be seen that the median scores of all signals differ with significance $p < 0.05$ in scenes 1 and 2. In acoustic scene 1, the median of the TFS method is 36.5, while for the TFSH method the median is much higher (80.5). In acoustic scene 2, the median of the TFS method is 55, while for the TFSH method this result is considerably higher (81). This clearly indicates that 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.

7 Discussion and Conclusion

In this paper two methods were compared in terms of (i) preserving spatial information and (ii) preserving the DOA of residual noise. The methods, referred to as TFS and TFSH, performed distinctly differently, where the TFS method outperformed the TFSH method for the former (desired signal spatial information preservation), and the TFSH method outperformed the TFS method for the latter (residual noise DOA preservation). In particular, regarding the preservation of desired signal spatial information, the performance of the TFSH

method is sensitive to source-microphone array separation. Further, with regard to residual noise DOA preservation, the TFS method is sensitive to the noise type. For highly reverberant environments with distant sources, the performance of the two methods is similar, and the TFSH method may be preferred for this case since it offers direct processing in the SH domain.

Acknowledgment

This research was supported by Facebook Reality Labs.

References

[1] Perry, T. S., "Virtual reality goes social," *IEEE Spectrum*, 53(1), pp. 56–57, 2016.

[2] Moore, C., "The Virtual Yellow House: Experimental tangling with virtual reality," *IEEE Consumer Electronics Magazine*, 5(4), pp. 103–104, 2016.

[3] Markwalter, B., "Entertainment and immersive content: What's in store for your viewing pleasure," *IEEE Consumer Electronics Magazine*, 4(1), pp. 83–86, 2015.

[4] Greenwald, S., Kulik, A., Kunert, A., Beck, S., Frohlich, B., Cobb, S., Parsons, S., Newbutt, N., Gouveia, C., Cook, C., et al., "Technology and applications for collaborative learning in virtual reality," 2017.

[5] Doclo, S., Gannot, S., Moonen, M., and Spriet, A., "Acoustic beamforming for hearing aid applications," *Handbook on array processing and sensor networks*, pp. 269–302, 2010.

[6] Doclo, S., Kellermann, W., Makino, S., and Nordholm, S. E., "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Processing Magazine*, 32(2), pp. 18–30, 2015.

[7] Hadad, E., Doclo, S., and Gannot, S., "The binaural LCMV beamformer and its performance analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3), pp. 543–558, 2016.

[8] Cornelis, B., Moonen, M., and Wouters, J., "Performance analysis of multichannel Wiener filter-based noise reduction in hearing aids under second order statistics estimation errors," *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5), pp. 1368–1381, 2010.

[9] Marquardt, D., Hohmann, V., and Doclo, S., "Interaural coherence preservation in multi-channel Wiener filtering-based noise reduction for binaural hearing aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(12), pp. 2162–2176, 2015.

[10] Klasen, T. J., Van den Bogaert, T., Moonen, M., and Wouters, J., "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," *IEEE Transactions on Signal Processing*, 55(4), pp. 1579–1585, 2007.

[11] Hadad, E., Marquardt, D., Doclo, S., and Gannot, S., "Extensions of the binaural MWF with interference reduction preserving the binaural cues of the interfering source," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 241–245, IEEE, 2016.

[12] Hadad, E., Marquardt, D., Doclo, S., and Gannot, S., "Extensions of the binaural MWF with interference reduction preserving the binaural cues of the interfering source," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 241–245, IEEE, 2016.

[13] Zohourian, M. and Martin, R., "GSC-based binaural speaker separation preserving spatial cues," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 516–520, IEEE, 2018.

[14] Enzner, G., Azarpour, M., and Siska, J., "Cue-preserving mmse filter for binaural speech enhancement," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5, IEEE, 2016.

[15] Rafaely, B., *Fundamentals of spherical array processing*, volume 8, Springer, 2015.

[16] Shabtai, N. R. and Rafaely, B., "Generalized spherical array beamforming for binaural speech

reproduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1), pp. 238–247, 2014.

[17] Sun, H., Yan, S., and Svensson, U. P., "Optimal higher order ambisonics encoding with predefined constraints," *IEEE transactions on audio, speech, and language processing*, 20(3), pp. 742–754, 2011.

[18] Borrelli, C., Canclini, A., Antonacci, F., Sarti, A., and Tubaro, S., "A Denoising Methodology for Higher Order Ambisonics Recordings," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 451–455, IEEE, 2018.

[19] Herzog, A. and Habets, E. A., "Direction Preserving Wiener Matrix Filtering for Ambisonic Input-output Systems," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 446–450, IEEE, 2019.

[20] Abend, U. and Rafaely, B., "Spatio-spectral masking for spherical array beamforming," in *Science of Electrical Engineering (ICSEE), IEEE International Conference on the*, pp. 1–5, IEEE, 2016.

[21] Moore, A. H., Lightburn, L., Xue, W., Naylor, P. A., and Brookes, M., "Binaural mask-informed speech enhancement for hearing aids with head tracking," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 461–465, IEEE, 2018.

[22] Lugasi, M. and Rafaely, B., "Speech Enhancement Using Masking for Binaural Reproduction of Ambisonics Signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.

[23] Rafaely, B. and Avni, A., "Interaural cross correlation in a sound field represented by spherical harmonics," *The Journal of the Acoustical Society of America*, 127(2), pp. 823–828, 2010.

[24] ITU-R, R., "1534-1,"Method for the subjective assessment of intermediate quality levels of coding systems (MUSHRA)"," *International Telecommunication Union*, 2003.

[25] Lindau, A., Erbes, V., Lepa, S., Maempel, H.-J., Brinkman, F., and Weinzierl, S., "A spatial audio quality inventory (SAQI)," *Acta Acustica united with Acustica*, 100(5), pp. 984–994, 2014.