BINAURAL SOUND REPRODUCTION BEAMFORMING USING SPHERICAL MICROPHONE ARRAYS

Noam R. Shabtai* *and Boaz Rafaely*†

Department of Electrical and Computer Engineering,
Ben-Gurion University of the Negev, Beer-Sheva, Israel.
*shabtai.noam@gmail.com †br@ee.bgu.ac.il

ABSTRACT

Currently employed microphone arrays usually have a single-channel output, such that no spatial information can be perceived by a human listener. However, spatial information may trigger spatial mechanisms in the human auditory system which can improve the intelligibility. This work presents a mathematical framework for the binaural beamforming approach for the ideal and order-limited representation in the spherical harmonics domain. The performance of the proposed binaural beamformer is compared to that of a monaural maximum directivity beamformer using objective signal-based measures and subjective listening tests. It is shown that using the binaural beamformer results in higher intelligibility than the monaural beamformer.

Index Terms— Beamforming, Binaural sound reproduction, Head-related transfer functions, Spherical Harmonics, Microphone arrays.

1. INTRODUCTION

The employment of microphone arrays in multi-participant telecommunication applications has become popular in recent years [1]. When microphone arrays are used, the recorded signals at all microphones can be combined such that a desired signal arrival direction can be emphasized. This procedure is referred to as beamforming. A number of approaches for beamforming are available, e.g., minimum variance distortionless response beamformer [2], generalized side-lobe canceller [3], etc. Using different combinations of individual microphone signals, a single microphone array can be used in order to generate several beams, one for each participant. Steering one of these beams towards the location of a given participant attenuates sounds originating from other directions. As a consequence, distracting sounds such as other talkers, reverberation, and background noise are attenuated. It has been shown that beamforming does indeed increase speech intelligibility in noisy and reverberant environments [4].

Additionally, signal processing methods such as blind source separation [5], echo cancellation [6], and dereverberation [7] can be employed in microphone arrays in order to increase intelligibility. However, beamforming systems which are currently employed in telecommunication applications typically produce a single-channel output, such that the inherent spatial information is limited. For that reason, the employment of conventional beamforming methods which are currently available, is more compatible with applications where the receiver is yet another machine.

In several studies, the improvement in intelligibility when binaural signals are used instead of monaural signals is investigated, and represented in the context of *spatial release from masking* (SRM) [8]. It has been shown that the human auditory system employs information in the binaural cues, *inter-aural time difference* (ITD) and *inter-aural level difference* (ILD), in order to achieve SRM [9]. In this context, the SRM was used to model the “cocktail party effect” [10]. Hence, recent investigated algorithms for binaural speech enhancement [11–13] and binaural dereverberation [14] are aiming towards the preservation of these binaural cues.

In this work, a binaural approach for beamforming is presented which is aiming at the preservation of the 3-D sound field in general, in addition to the information which is contained in the binaural cues. A mathematical framework was recently presented for the binaural beamforming approach [15] in which spatial sound reproduction is incorporated in the beamforming process using the *head related transfer functions* (HRTFs) [16]. This work extends [15] by presenting a discussion on the order selection of the weight function in the spherical harmonics domain, when the order of the estimated plane-wave amplitude-density function is limited by the array. Furthermore, in addition to a more comprehensive objective signal-based analysis, the performance of the proposed binaural beamformer is compared to that of a monaural maximum directivity beamformer using a subjective listening test. Percentage correct results which are averaged over 5 subjects show that the binaural beamformer outperforms the monaural beamformer by 33.3% correct decisions for an input *signal-to-noise ratio* (SNR) of -40dB.

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

2. BINAURAL BEAMFORMING

In conventional beamforming, a signal direction may be emphasized using a weighted sum of the microphone signals to form a single-channel output signal

$$y\left(k\right)=\sum_{q=1}^{Q}w_{q}^{\star}\left(k\right)p_{q}(k) \tag{1}$$

where $w_q\left(k\right)$ represents the beamformer q'th weight at wave number k. At this wave number, the recorded signal at the q'th microphone is denoted by $p_q\left(k\right)$.

The pressure level at the origin of the array may be represented using an integral over a continuum of plane-waves

$$p_0\left(k\right)=\int_{\Omega}a\left(k,\Omega\right)\mathrm{d}\Omega \tag{2}$$

where $a\left(k,\Omega\right)$ is a plane-wave amplitude spherical-density function, at the wave number k and at the arrival spatial angle of $\Omega=\left(\theta,\phi\right)$ in a spherical coordinate system such that

$$\mathrm{d}\Omega=\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi \tag{3}$$

The weights may be calculated such that an optimal directivity criterion is obtained typically subject to a distortionless response constraint, of which procedure is referred to in this work as monaural beamforming.

The proposed *binaural* beamforming approach, shown in Fig. 1, incorporates spatial sound reproduction using the HRTFs in the beamforming procedure [15]. This is essentially different than filtering the output of a monaural beamformer with the HRTF, which does not reflect the spatial information in the recording venue, and therefore does not maintain a spatial discrimination between the target signal and the noise. A preliminary step of *plane wave decomposition* (PWD) is performed in order to estimate the plane-wave amplitude density function [17–19]. Then the output of the beamformer is calculated using

$$\begin{aligned} y_L\left(k\right) &= \int_{\Omega\in S^2}w^{\star}\left(\Omega\right)a\left(k,\Omega\right)H_L\left(k,\Omega\right)\mathrm{d}\Omega && (4)\\ y_R\left(k\right) &= \int_{\Omega\in S^2}w^{\star}\left(\Omega\right)a\left(k,\Omega\right)H_R\left(k,\Omega\right)\mathrm{d}\Omega && (5)\end{aligned}$$

where H_L and H_R are the HRTF of the left and right ears, respectively. Then, using headphones, the beamformer outputs y_L and y_R may be played at the left and right ears, respectively. The weights of this binaural beamformer are calculated in order to obtain a maximum directivity for a slow-varying HRTF function of Ω.

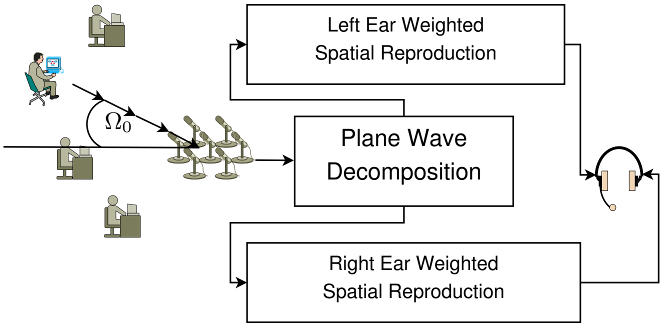

Fig. 1*:* Proposed binaural beamforming approach.

3. IDEAL MATHEMATICAL FRAMEWORK

In this section, a mathematical formulation of the proposed approach is presented in the case where the spherical Fourier representation of y, w, a, and H is not order limited. Hence

$$a\left(k,\Omega_k\right)=\sum_{n=0}^{\infty}\sum_{m=-n}^{n}a_{nm}\left(k\right)Y_n^m\left(\Omega_k\right) \tag{6}$$

where $a_{nm}\left(k\right)$ is the spherical Fourier transform of $a\left(k,\cdot\right)$ which equals to

$$a_{nm}\left(k\right)=\int_{\Omega\in S^2}a\left(k,\Omega\right)Y_n^{m\star}\left(\Omega\right)\mathrm{d}\Omega \tag{7}$$

The term Y_n^m is the mn'th spherical harmonic

$$Y_n^m\left(\theta,\phi\right)\triangleq\sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}}P_n^m\left(\cos\left(\theta\right)\right)e^{im\phi} \tag{8}$$

where $P_n^m\left(\cdot\right)$ is the associated Legendre function. The indices nm maintain $n\in\mathbb{N}$ and $m\in\{-n\dots n\}\,\forall n$.

In the case of a single unit amplitude plane wave arriving from Ω_0, the output of the beamformer in Eqs. (4) and (5) reduces to

$$y\left(k\right)=w^{\star}\left(\Omega_0\right)H\left(k,\Omega_0\right) \tag{9}$$

where the indication for left and right ears was omitted for convenience. Hence w may be regarded as a function which spatially emphasizing arrival directions, and the frequency response to a plane wave is proportional to the HRTF in the arrival direction.

In general however, the output of the beamformer may be represented in the spherical harmonics domain by applying Parseval's theorem on Eqs. (4) and (5)

$$y\left(k\right)=\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\tilde{w}_{nm}^{\star}\left(k\right)a_{nm}\left(k\right) \tag{10}$$

The term $\tilde{w}_{nm}\left(k\right)$ is the spherical Fourier transform of $wH^{\star}\left(k,\cdot\right)$ which equals to

$$\tilde{w}_{nm}\left(k\right)=\int_{\Omega\in S^2}w\left(\Omega\right)H^{\star}\left(k,\Omega\right)Y_n^{m\star}\left(\Omega\right)\mathrm{d}\Omega \tag{11}$$

Substituting spherical Fourier transform of w and H in Eq. (11) yields

$$\begin{aligned}\tilde{w}_{nm}(k) &= \sum_{n_1=0}^{\infty}\sum_{m_1=-n_1}^{n_1} w_{n_1m_1}\sum_{n_2=0}^{\infty}\sum_{m_2=-n_2}^{n_2} H^{\star}_{n_2m_2}(k) \\ &\quad \int_{\Omega\in S^2} Y_{n_1}^{m_1}(\Omega)\, Y_{n_2}^{m_2\star}(\Omega)\, Y_n^{m\star}(\Omega)\, \mathrm{d}\Omega \end{aligned} \tag{12}$$

Solution to the integral over three spherical harmonics is given by [20]

$$\begin{aligned} A_{n_1n_2n}^{m_1m_2m} &\triangleq \int_{\Omega\in S^2} Y_{n_1}^{m_1}(\Omega)\, Y_{n_2}^{m_2\star}(\Omega)\, Y_n^{m\star}(\Omega)\, \mathrm{d}\Omega \\ &= \sqrt{\frac{(2n+1)(2n_2+1)}{4\pi(2n_1+1)}} \times \\ &\quad \mathcal{C}(nn_2n_1;000)\,\mathcal{C}(nn_2n_1;mm_2m_1) \end{aligned} \tag{13}$$

where $\mathcal{C}$ are the Clebsch Gordan coefficients [21] which may be expressed using Wigner's closed-form [22] or in the Racha's closed-form [23].

4. ORDER-LIMITED REPRESENTATION

In practice, $a(k,\Omega)$ can be estimated up to a certain order N_a which is dependent on the number and location of microphones. That is, only $a_{00}(k)\ldots a_{N_aN_a}(k)$ can be estimated. The calculation of $a_{nm}(k)$ is performed by capturing the pressure at S points on a sphere to yield $p(k,\Omega_s)\,, \quad s = 1\ldots S$. Finally, $a_{nm}(k)$ are found by applying

$$a_{nm}(k) \approx \frac{p_{nm}(k)}{b_n(kr)} \quad , n = 1\ldots N_a \quad , m = -n\ldots n \tag{14}$$

where it is assumed that the microphones are located on a rigid sphere, and r is the radius of the spherical array. The term $b_n(kr)$ satisfies [24]

$$b_n(kr) = 4\pi i^n\left[j_n(kr) - \frac{j_n'(kr)}{h_n'^{(2)}(kr)}h_n^{(2)}(kr)\right] \tag{15}$$

where $j_n(\cdot)$ and $j_n'(\cdot)$ are the spherical Bessel function of the first kind and its derivative, respectively. Also, $h_n^{(2)}(\cdot)$ and $h_n'^{(2)}(\cdot)$ are the spherical Hankel function of the second kind and its derivative, respectively.

Denoting the order of w and H with N_w and N_H, respectively, allows to formulate the order of $\tilde{w}$, $N_{\tilde{w}}$. The order of multiplication of polynomials equals to the sum of the polynomial orders. Since the spherical harmonics representation of w and H use polynomials in $\cos\theta$, $N_{\tilde{w}} = N_w + N_H$. The output of the beamformer is therefore set by

$$y(k) = \sum_{n=0}^{\min\{N_a, N_w+N_H\}} \sum_{m=-n}^{n} \tilde{w}^{\star}_{nm}(k)\, a_{nm}(k) \tag{16}$$

Now, the order of a is limited by the array. The order of w is arbitrary. The order of H may be limited by the sampling scheme used in the generation of the HRTF database. However, since the measuring of the HRTF may use moving microphones it is essentially not limited, so the order of H may be considered arbitrary as well. Hence, if we select the order of w, the orders of H which are higher than $N_a - N_w$ are redundant, and vice versa, if we select the order of H, the orders of w which are higher than $N_a - N_H$ are redundant.

5. EVALUATION BY SIMULATION

The proposed binaural beamformer was simulated with a look direction in the front of the head at $\Omega_l = (90°, 90°)$. The CIPIC database of HRTFs [25] that were measured in an anechoic setting was used, and their spherical Fourier transforms were calculated to the order of $N_H = 10$. The weight function was selected such that

$$w_{nm} = Y_n^{m\star}(\Omega_l) \tag{17}$$

where $n \leq N_w = 5$.

Figure 2 shows the one-dimensional beam pattern of the binaural beamformer at 1KHz, where the arrival elevation is fixed at $\theta_0 = 90°$, and the arrival azimuth is a parameter $\phi \in [0, 180°]$. The output of the binaural beamformer at each ear is compared to the output of a monaural maximum-directivity beamformer which is calculated using $y(k) = \sum_{n=0}^{N_w}\sum_{m=-n}^{n} Y_n^m(\Omega_l)\, a_{nm}(k)$ [18], and therefore equal at both ears. It may be seen that using either one of the beamformers, the signal at the look direction is enhanced. Furthermore, the head symmetry can be noticed between the left and right ears with respect to the front of the head. Also, both binaural and monaural beamformers are similar in terms of spatial directivity, so any differences in perception cannot be attributed to spatial attenuation.

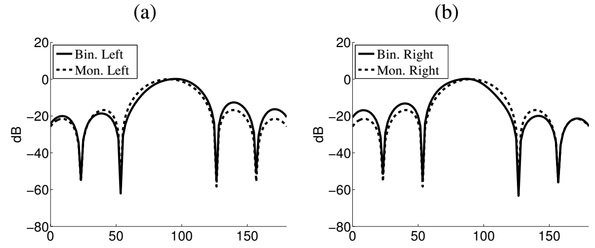

Fig. 2*:* Beam pattern of the binaural beamformer at 1KHz is displayed with solid curve, for (a) left ear and (b) right ear. Beam pattern of a monaural maximum-directivity beamformer is displayed with a dashed curve.

The binaural beamformer was simulated again with a signal arriving either from the look direction, i.e. where $\Omega_0 = (90°, 90°)$, or from the interferer direction, i.e. where

$\Omega_0 = (90°, 0°)$. Figure 3 displays the response of the binaural beamformer vs. the original HRTF in the two cases at each ear. The similarity of the binaural beamformer response to the HRTF can clearly be noticed when the target signal is in the look direction. Moreover, in both directions, the output of the binaural beamformer is similar to the HRTF, therefore facilitating binaural sound reproduction. However, when the target is located 90° away in azimuth from the look direction, the response is attenuated from the HRTF, which is the result of the weight function.

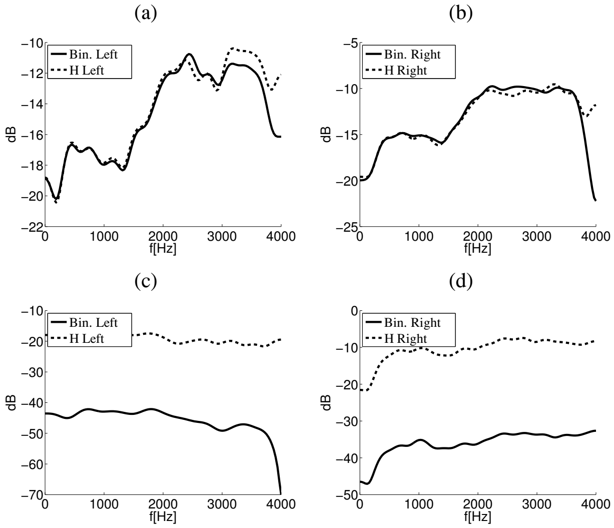

Fig. 3*:* Frequency response of the binaural beamformer (solid) compared with HRTF (dashed) where look direction is $\Omega_l = (90°, 90°)$. Response shown at (a) left ear with $\Omega_0\,(90°, 90°)$, (b) right ear with $\Omega_0\,(90°, 90°)$, (c) left ear with $\Omega_0\,(90°, 0°)$, (d) right ear with $\Omega_0\,(90°, 0°)$.

6. EVALUATION USING LISTENING TEST

Subjective listening test was performed in order to evaluate the performance of the binaural beamformer in terms of enhancement in intelligibility when compared to the monaural beamformer. In this test only anechoic speech was used, and reverberation was not simulated. Speech segments of two male and two female speakers from the *coordinate response measure* (CRM) speech identification test [26] were employed. This database consists of sentences with the following structure "ready CALLSIGN, go to COLOR NUMBER now!". Colors are: "Blue", "Red", "White", and "Green". Numbers are "One" to "Eight", and CALLSIGN is a person's name. The aim of the test is to identify the color and number for a specific pre-designed call sign. In the original simulated sound field, a target speaker was placed in front of the head at $\Omega_T = (90°, 90°)$, and an interfering speaker was placed at the right ear direction, $\Omega_I = (90°, 0°)$. The experimental setup was composed of a personal computer and a computer screen to display the experiment information. A Keyboard was used for subject response and feedback. The sounds were played using AKG K702 headphones. In this experiment, the interfering speaker always used "Blue" and "Seven" for color and number, respectively. The subjects were instructed to identify the color and the number of the target speaker, knowing that it can neither be "Blue" nor "Seven". Target and interfering speakers were always from the same gender. The call signs of the male interferer and target speakers were "Charlie" and "Ringo", respectively. The call signs of the female interferer and target speakers were "Charlie" and "Tiger", respectively. The test material consisted of 21 sentences, delivered at an SNR of -40dB at the array input. Table 1 shows percentage-correct results of 5 subjects, which were students in the research lab. Although a more comprehensive experiment with more subjects and a wider SNR range is necessary, this feasibility experiment shows that better performance is achieved by the proposed binaural beamforming system.

Table 1*:* Percentage correct decision of 5 subjects in CRM listening test when the binaural and monaural beamformers are used, for -40dB SNR at array input.

Subject	1	2	3	4	5
Binaural	75.0	83.3	83.3	91.7	75.0
Monaural	41.7	33.3	58.3	75.0	33.3
Diff.	33.3	50.0	25.0	16.7	41.7
Av. Diff.	33.3				

7. CONCLUSION

A binaural beamforming approach was proposed which incorporates the HRTF in a spatial reproduction scheme. Beam patterns were displayed for the left and right ears, compared with a beam pattern of a monaural maximum directivity beamformer. The frequency response of the binaural beamformer retrieves the HRTF at the look direction, and proportionally attenuates the HRTF in the direction of the noise. A preliminary listening test with 5 subjects using the CRM corpus has shown that the intelligibility is improved by an average of 33.3% in terms of percentage correct decision, when using the binaural beamformer.

8. REFERENCES

[1] J. Flanagan, J. Johnston, R. Zahn, and G. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Am.*, vol. 78, no. 5, pp. 1508–1518, Apr. 1985.

[2] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 10, pp. 1365–1376, Oct. 1987.

[3] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[4] G. Saunders and J. Kates, "Speech intelligibility enhancement using hearing-aid array processing," *J. Acoust. Soc. Am.*, vol. 102, no. 3, pp. 1827–1837, Sept. 1997.

[5] R. Aichner, H. Buchner, F. Yan, and W. Kellermann, "A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments," *Signal Processing*, vol. 86, pp. 1260–1277, Oct. 2006.

[6] F. Küch and W. Kellermann, "Orthogonalized power filters for nonlinear acoustic echo cancellation," *Signal Processing*, vol. 86, pp. 1168–1181, June 2006.

[7] H. Buchner and W. Kellermann, "TRINICON for dereverberation of speech and audio signals," in *Speech Dereverberation*, P.A. Naylor and N.D. Gaubitch, Eds., pp. 311–385. Springer-Verlag, London, 2010.

[8] R. L. Martin, K. I. McAnally, R. S. Bolia, G. Eberle, and D. S. Brungart, "Spatial release from speech-on-speech masking in the median sagittal plane," *J. Acoust. Soc. Am.*, vol. 131, no. 1, pp. 378–385, Jan. 2011.

[9] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, pp. 233–237, May 2012.

[10] G. L. Jones and R. Y. Litovsky, "A cocktail party model of spatial release from masking by both noise and speech interferers," *J. Acoust. Soc. Am.*, vol. 130, no. 3, pp. 1463–1474, Sept. 2011.

[11] T. Van den Bogaert, S. Doclo, M. Moonen, and J. Wouters, "The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids," *J. Acoust. Soc. Am.*, vol. 124, no. 1, pp. 484–497, 2008.

[12] S. Doclo, M. Moonen, T. Van den Bogaert, and J. Wouters, "Reduced-bandwidth and distributed mwf-based noise reduction algorithms for binaural hearing aids," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 1, pp. 38–51, 2009.

[13] A. Bertrand and M. Moonen, "Distributed adaptive node-specific signal estimation in fully connected sensor networks - part i: sequential node updating," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5277–5291, 2010.

[14] M. Jeub, M. Schäfer, T. Esch, and P. Vary, "Model-based dereverberation preserving binaural cues," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 7, pp. 1732–1745, 2010.

[15] N. R. Shabtai and B. Rafaely, "Spherical array beamforming for binaural sound reproduction," in *Proc. 27th IEEE Convention of Electrical and Electronics Engineers in Israel (IEEEI 2012)*, Eilat, Israel, 2012, pp. 1–5.

[16] D. S. Brungart and W. M. Rabinowitz, "Auditory localization of nearby sources. head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1465–1479, Sept. 1999.

[17] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Proc. 2002 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2002)*, Orlando, FL, USA, 2002, pp. 1781–1784.

[18] B. Rafaely, "Plane-wave decomposition of the sound field on a sphere by spherical convolution," *J. Acoust. Soc. Am.*, vol. 116, no. 4, pp. 2149–2157, Oct. 2004.

[19] D. Zotkin, R. Duraiswami, and N. Gumerov, "Plane-wave decomposition of acoustical scenes via spherical and cylindrical microphone arrays," *IEEE Trans. on Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 2–16, Jan. 2010.

[20] W. J. Thompson, *Angular Momentum: An Illustrated Guide to Rational Symmetries for Physical System*, chapter Appendix III, p. 428, Wiley-VCH Verlag GmbH & Co., 2004.

[21] M. E. Rose, *Elementary theory of Angular Momentum*, chapter 3, pp. 39–40, John Wiley & Sons Inc., New York, 1959.

[22] E. P. Wigner, *Grouppentheorie*, Friedrich, Vieweg und Sohn, Braunshweig, Germany, 1931.

[23] G. Racah, "Theory of complex spectra. II," *Phys. Rev.*, , no. 62, pp. 438–462, Nov. 1942.

[24] G. Arfken and H. J. Webber, *Mathematical methods for physicists*, chapter 11, pp. 669–722, Academic Press, San Diego, 5 edition, 2001.

[25] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. 2001 IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA 2001)*, New Paltz, NY, USA, 2001, pp. 99–102.

[26] R. S. Bolia, W. T. Nelson, M. A. Ericson, and B. D. Simpson, "A speech corpus for multitalker communications research," *J. Acoust. Soc. Am.*, vol. 107, no. 2, pp. 1065–1066, Feb. 2000.