Mixing time prediction using spherical microphone arrays

Philipp Götz, Konrad Kowalczyk,[a] and Andreas Silzle
Fraunhofer Institute for Integrated Circuits IIS, Audio & Multimedia Division, 91058 Erlangen, Germany
philipp.goetz@iis.fraunhofer.de, konrad.kowalczyk@agh.edu.pl, andreas.silzle@iis.fraunhofer.de

Emanuël A. P. Habets
International Audio Laboratories Erlangen, 91058 Erlangen, Germany
emanuel.habets@audiolabs-erlangen.de

Abstract: Human perception of room acoustics depends among others on the time of transition from early reflections to late reverberation in room impulse responses, which is known as mixing time. In this letter, a multi-channel mixing time prediction method is proposed, which in contrast to state-of-the-art channel-based predictors accounts for spatio-temporal properties of the sound field. The proposed diffuseness-based method is compared with existing model- and channel-based prediction methods through measurements and acoustic simulations, and is shown to correlate well with the perceptual mixing time. Furthermore, insights into relations between prediction methods and mixing time definitions based on reflection density are presented.

[NX]
Date Received: July 31, 2014 **Date Accepted:** December 18, 2014

1. Introduction

Reflection of incident sound at boundary surfaces causes the impulse-excited sound field in an enclosure to become increasingly diffuse as time progresses. The moment of transition from coherent early reflections to diffuse late reverberation in room impulse responses (RIRs)—referred to as mixing time t_m—plays an important role in psychoacoustics.[1] Mixing time represents a perceptual phenomenon caused by the limited temporal and spatial resolution of the human auditory system, in particular, its poor sensitivity toward interaural time and level differences in an increasingly diffuse sound field, which results in the loss of directional information.[2] In general, mixing time prediction methods can be categorized into model- and channel-based methods.

Empirically derived model-based methods predict the mixing time based on global characteristics of an enclosure such as volume (Vol),[3] boundary surface area,[4] reverberation time (RT),[5] or mean free-path length (MFPL).[6] Alternatively, several methods define the mixing time as a moment in a RIR where the number of reflections within a window of predefined duration reaches a specific threshold. In literature, these empirically found reflection density numbers range from 250 (Ref. 7) through 1000 (Ref. 8) and 2000 (Ref. 9), up to 10 000 (Ref. 10) reflections per second. According to Polack,[4] 10 incoherent reflections within a window of 24 ms suffice to reach a fully mixed sound field.

State-of-the-art channel-based prediction methods utilize statistical analysis from measured RIRs and define a mixing time criterion (with a specific threshold value) which is derived from an underlying physical assumption about the sound field.

[a]Author to whom correspondence should be addressed. Present address: AGH University of Science and Technology, 30-059 Krakow, Poland.

The methods proposed by Abel and Huang[11] and Stewart and Sandler[12] assume that the amplitudes of diffuse late reverberation resemble a Gaussian noise with frequency-dependent coloring and decay rate. On the other hand, the method proposed by Hidaka *et al.*[13] is based on the decreasing correlation between the time-frequency distribution of acoustic energy of the entire impulse response and that of the latter windowed RIR segments. Yet another prediction method proposed by Defrance *et al.*[14] defines the mixing time as the moment after which the number of cumulated arriving reflections can be described by a linearly increasing function. The similarity of the mixing time predicted using these methods and the perceptual mixing time, defined as the instant after which directional information is no longer perceived, was investigated by Lindau *et al.*[15]

Existing channel-based methods are applied to a single RIR and hence exploit solely the temporal characteristics of the sound field. As pointed out by Lindau *et al.*,[15] the strong positional variance of channel-based methods necessitates the averaging of mixing times over multiple measurement positions in an enclosure in order to obtain a reliable estimate. The predictor proposed in this work exploits the spatiotemporal information by performing an analysis of the diffuseness of the sound field[1] computed from the RIRs measured using a microphone array. As shown in this letter, this multi-channel prediction method leads to the results that are closer to the perceptual mixing time than it is the case for single-channel predictors, and, furthermore, a reduction in positional variance compared to single-channel methods can be achieved. Some preliminary results have been shown in Ref. 16, however, here a thorough investigation is performed based on experiments in several different enclosures and multiple measurement positions, and the proposed method parametrization is argued on the basis of numerical simulations and consistency with perception. In addition, new insights are presented by relating the model- and channel-based predictors to the methods based on the reflection density. This letter is structured as follows. Two existing channel-based prediction methods that yield close results to the perceptual mixing time identified by Lindau *et al.*[15] are reviewed in Sec. 2 and the proposed approach is described in detail in Sec. 3, respectively. The experimental results are presented in Sec. 4, followed by concluding remarks in Sec. 5.

2. Physical and perceptual mixing time prediction

In this section, two channel-based prediction methods are briefly reviewed and the main conclusions of Lindau *et al.*[15] are summarized. Abel and Huang[11] introduced the so-called *normalized echo density profile* (NEDP), which describes the degree of Gaussianity of the probability density function of amplitudes in consecutive analysis frames. The authors used an analysis frame length of 23 ms, and the mixing time was found as the instant when the NEDP, denoted by $\eta(t)$, reaches the value of 1 for the first time. On the other hand, Hidaka *et al.*[13] presented a method which exploits the frequency-dependent attenuation of sound reflected at boundaries. To find t_m, a spectral correlation function $\rho(t)$ was introduced, which is defined as the Pearson correlation coefficient between the magnitude spectrum of the entire RIR and the magnitude spectrum of the RIR segment that begins at consecutive time instances t and finishes at the RIR end (cf. Schröder backwards integration). The mixing time is found when the spectral correlation averaged over the frequencies 353 Hz to 2.8 kHz falls below the value of $1/e$ for the first time. As illustrated in Fig. 1, the mixing times estimated using both methods are characterized by high positional variance, in contrast to the diffuseness-based method proposed in Sec. 3.

Many model- and channel-based predictors were perceptually evaluated by Lindau *et al.*[15] in binaural listening experiments. It was concluded that (1) predictions after Abel and Huang[11] are the closest to the perceptual mixing time among channel-based approaches and (2) model-based mixing time prediction is inconsistent with perception in very small and very large enclosures. Finally, based on regression analysis of the listening test results, Lindau *et al.*[15] introduced the perceptual mixing time $t_{m,p}$ for the 50th and 95th percentiles, which can be computed based on prediction according to Abel and Huang.[11]

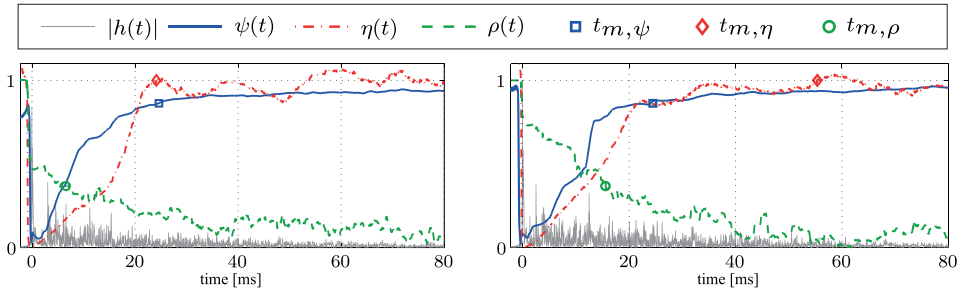

Fig. 1. (Color online) Mixing time prediction for two different measurement positions (i.e., loudspeaker and microphone positions) in a room shown in the left and right figures: Absolute value of the impulse response (solid gray line), diffuseness (solid black line), NEDP $\eta(t)$ (dashed-dotted line), and spectral correlation function $\rho(t)$ (dashed line). The estimated mixing times t_m are denoted by the square, diamond, and circle markers, respectively.

3. Proposed diffuseness-based method

For robust mixing time prediction, a spatiotemporal measure could be advantageous. In this letter, we propose to use a physical quantity that measures the process of sound field diffusion in an enclosure known as diffuseness. In acoustic spaces, diffuseness follows a gradual change from an initial concentration of acoustic energy toward its homogeneous distribution and analogously, the gradual change from an initially directional energy transport toward its isotropy. Since such a measure requires three-dimensional sound field analysis, a spherical microphone array can conveniently be applied. The diffuseness $(0 \leq \hat{\psi}(k,t) \leq 1)$ can be approximated by the temporal variation of sound intensity using[17,18]

$$\hat{\psi}(k,t) = \sqrt{1 - \frac{\|\mathrm{E}\{\mathbf{I}(k,t)\}\|}{\mathrm{E}\{\|\mathbf{I}(k,t)\|\}}}, \tag{1}$$

where $\mathbf{I}(k,t)$ denotes the sound intensity vector, $\mathrm{E}\{\cdot\}$ denotes the expectation operation (i.e., time averaging), $\|\cdot\|$ is the l^2-norm of a vector, t denotes time, and $k = 2\pi f/c$ denotes the wave number, f is the frequency, and c denotes the propagation speed. The sound intensity $\mathbf{I} = \frac{1}{2}\Re\{p^* \cdot \mathbf{v}\}$ is computed from the sound pressure p and particle velocity vector $\mathbf{v}$. Measuring RIRs with a spherical microphone array, the sound intensity can be conveniently computed in the spherical harmonic domain. The spherical harmonic decomposition of the measured RIRs can be performed using[19]

$$H_{lm}(k,t,r) = \int_0^{2\pi}\int_0^{\pi} H_w(k,t,r,\theta,\phi)\, Y_{lm}^*(\theta,\phi)\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi, \tag{2}$$

where $\theta \in [0,\pi]$ and $\phi \in [0,2\pi]$ are the elevation and azimuth angles, respectively, r is the array radius, and $Y_{lm}(\theta,\phi)$ denotes the spherical harmonics of order l and degree m. The transfer function of the windowed RIR measured at position (r,θ,ϕ) with the window centered at time t is denoted as H_w. The sound pressure can then be calculated from the omnidirectional zero-order spherical harmonic, while the particle velocities $\mathbf{v} = [v_x, v_y, v_z]^{\mathrm{T}}$ along Cartesian coordinate axis are obtained by a linear combination of the rotated first-order spherical harmonics, which is given by[20]

$$p(k,t) = \frac{H_{00}(k,t)}{b_0(k)}, \quad v_i(k,t) = \frac{1}{b_1(k)}\sum_{m=-1}^{1}\beta_{i,m}H_{1,m}(k,t), \tag{3}$$

where $i \in \{x,y,z\}$, the normalization factors $1/b_l(k)$ are the modal radial filters compensating for the mode amplification,[21] and $\beta_{x,m} = Y_{lm}(\pi/2,\pi)$, $\beta_{y,m} = Y_{lm}(\pi/2,-\pi/2)$, and $\beta_{z,m} = Y_{lm}(\pi,0)$, respectively. Next the broadband diffuseness measure is computed

from narrowband estimates by frequency averaging over the range $[k_{\min}, k_{\max}]$. Based on broadband diffuseness, we can compute an estimate of the broadband direct-to-diffuse ratio (DDR) $\hat{\Gamma}(t)$, which describes the ratio between direct and diffuse energy:[22]

$$\hat{\Gamma}(t) = 10\log_{10}\left(\frac{1}{\hat{\psi}(t)} - 1\right) \quad [\text{dB}]. \tag{4}$$

Finally, the predicted mixing time is found as the moment when the broadband DDR exceeds a specific threshold value (as given in Sec. 4).

4. Experimental evaluation

4.1 Parametrization and setup

Since the measure of diffuseness relies on directional variation of sound intensity over time, selecting the appropriate analysis windows (i.e., an averaging period for diffuseness estimation, and the frame length and the hop size of consecutive frames for intensity analysis) is very important for reliable mixing time prediction. In order to obtain sufficiently many intensity values within a longer diffuseness analysis window with good time-frequency resolution, intensity vectors were estimated in short frames of 64 samples (1.5 ms), which enables capturing of a single reflection, and a hop size of 1 sample was chosen. Note that this short window length is determined independent of the room size. The diffuseness was then estimated using recursive averaging with time constant $\tau = 12$ ms to ensure sufficiently high temporal responsiveness to the changing sound field diffuseness while still allowing to estimate low DDRs. The final broadband diffuseness value was computed by averaging the narrowband estimates within the frequency range [1, 4.5] kHz in order to avoid noise amplification at low frequencies and spatial aliasing at high frequencies for the specific microphone array used for measuring RIRs. The mixing time was found when the DDR threshold of -8 dB was reached for the first time for the above analysis parametrization. Note that the selected analysis parametrization has been found to predict most consistently the mixing time (located between the 50th and 95th percentiles of the perceptual mixing time obtained by Lindau *et al.*[15]) in various enclosures.

To further motivate the parameter choice, the effect of using a larger hop size (50% overlap) or higher time averaging constant (50 ms) is shown in Fig. 2, which depicts the estimated diffuseness for a sudden transition from a perfectly non-diffuse (i.e., one plane wave) to an ideally diffuse sound field over time [Fig. 2 (left)] and the estimated DDRs for different simulated DDR values [Fig. 2 (right)]. Ideal non-diffuse and diffuse sound field conditions were created by synthesizing a single plane wave and by a superposition of 500 synthesized unit-amplitude plane waves arriving with random phase from directions uniformly distributed on a sphere, respectively. Using 50% overlap, high diffuseness (low DDR) values cannot be attained even in ideally diffuse fields. On the other hand, the time averaging of 50 ms leads to slow responsiveness

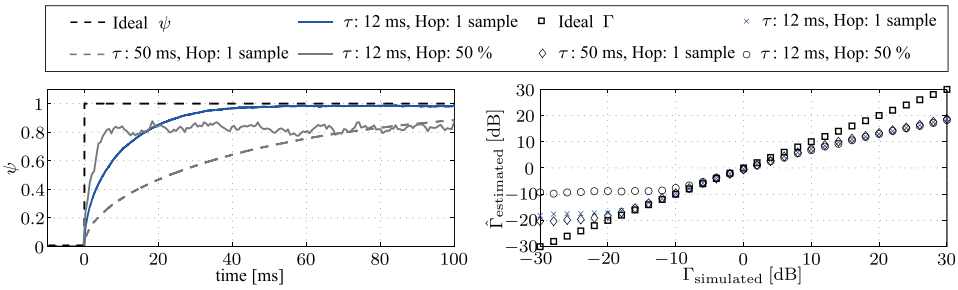

Fig. 2. (Color online) Left: A sudden transition from the synthesized ideal direct to the ideal diffuse field as a function of time. Right: Estimated steady-state DDRs as a function of simulated ideal sound fields.

to the changes in the sound field and the advantage of estimating lower DDR values is very minor [see Fig. 2 (right)].

For the proposed method, the RIR measurements were performed using the Eigenmike® spherical microphone array of *mhacoustics*,[23] while a high-quality omnidirectional microphone located at the same position was applied for single-channel methods. The three rooms under investigation include: (A) a small room ($T_{60} = 0.32$ s, floor area of 35 m^2, and volume of 105 m^3), (B) a medium size room ($T_{60} = 0.49$ s, floor area of 90 m^2, and volume of 495 m^3), and (C) a large hall ($T_{60} = 1.89$ s, floor area of 412 m^2, and volume of 3584 m^3), respectively. Note that all rooms were nearly empty during the measurements. Eight different transmission paths (i.e., source-receiver positions) were measured in each room using a logarithmic sine-sweep excitation; the sampling frequency was 44.1 kHz.

4.2 Mixing time from measurements

In Fig. 3, the mixing time results of the proposed method are compared to the state-of-the-art predictors of Abel and Huang,[11] two perceptual definitions proposed by Lindau *et al.*[15] (denoted as Lindau_{50} and Lindau_{95}), and the method after Hidaka *et al.*[13] Additionally, predictions according to three model-based methods depending on Vol,[3] RT,[5] and MFPL[6] are shown for comparison (an overview of these methods is provided, e.g., by Lindau *et al.*[15]). Note that for Room C, only one model-based prediction result (that is based on RT) is depicted as the geometry of the large hall was too complex for a reliable estimate of its volume and mean free-path length.

As can be observed, the proposed method correlates well with perception since the predicted mixing times are consistently lying in between the 50th and 95th percentiles defined after Lindau *et al.*,[15] while this is not the case for state-of-the-art approaches. Furthermore, the proposed method exhibits significantly less positional variance across different locations in an enclosure, which can effectively reduce the measurement and analysis effort. In Rooms A and B, the method based on RT also yields good results but it fails completely in a large enclosure such as Room C.

4.3 Mixing time vs reflection density

In this section, the relations between the mixing times predicted using the presented channel- and model-based methods and those based on reflection density are indicated. For this purpose, image-source models for Rooms A and B were created and all eight measurement positions were simulated. Room C was excluded from simulations due to its complex geometry. The number of reflections arriving in a period of 24 ms (following the definition by Polack[4]) was computed for the mixing times estimated with all discussed methods and for all measurement positions. The variation in the number of reflections for the three model-based predictors was due to the different reflection densities for the eight measurement positions at the predicted mixing time. The values

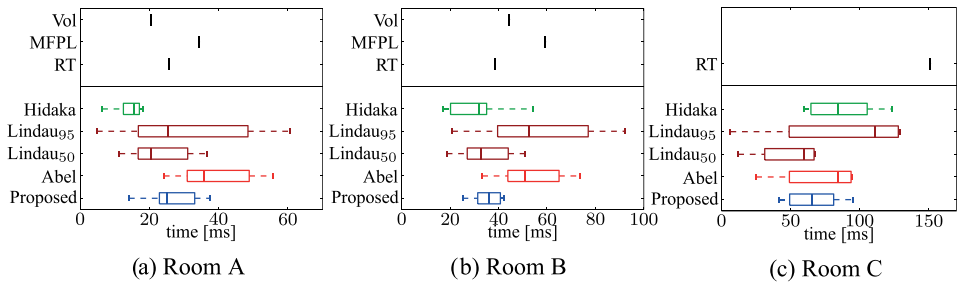

Fig. 3. (Color online) The median, inter-quartile range, and inner fence for the mixing times estimated using channel-based methods across eight measurement positions in each of the three rooms. Single-value results for model-based methods that depend on the RT, MFPL, and Vol are also shown.

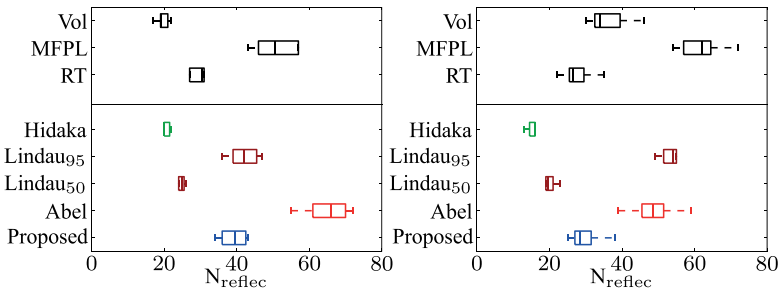

Fig. 4. (Color online) The median, inter-quartile range, and inner fence of the number of reflections arriving within a window of 24 ms starting at the estimated mixing time for eight measurement positions in Rooms A and B.

originally provided in seconds for methods[7–10] were recalculated for a 24 ms window for easier comparison.

As depicted in Fig. 4, remarkably, all methods exceed the well-known mixing time definition of Polack,[4] i.e., 10 incident reflections; also the perceptual mixing time after Lindau *et al.*[15] always yields a higher reflection number. The results obtained with all predictors are also significantly different from the values found in Refs. 7 and 10, where reflection densities of 6 and 240 are given, respectively. On the other hand, the reflection density corresponding to the mixing time predicted with the proposed approach, which ranges from 30 to 40 reflections arriving within a 24 ms period, is lying in between the values given by Schroeder[8] and Schreiber,[9] i.e., 24 and 48, respectively. The method of Hidaka *et al.*[13] yields 15–20 reflections, while that of Abel and Huang[11] corresponds to 45–65 reflections for considered rooms. Comparing the prediction results with perceptual values, it may be argued that mixing occurs at an instant when a higher number of reflections arrive within a 24 ms period than that given by Polack.[4]

5. Conclusion

In contrast to the state-of-the-art methods, the proposed mixing time prediction method is based on a spatiotemporal measure of diffuseness, which is advantageous since spatial distribution of the sound field is taken into account. The analysis parametrization and the mixing threshold are empirically found such that high consistency with perception is attained. Furthermore, the proposed method improves positional invariance thereby reducing the measurement and analysis effort. Finally, the relations between the mixing times predicted using existing and proposed methods and approaches utilizing reflection density are highlighted.

References and links

[1]H. Kuttruff, *Room Acoustics* (CRC Press, London, 2009).
[2]J. Blauert, *Spatial Hearing* (MIT Press, Harvard, MA, 1997).
[3]J.-M. Jot, L. Cerveau, and O. Warusfel, "Analysis and synthesis of room reverberation based on a statistical time-frequency model," in *Proceedings of the 103rd AES Convention*, New York (1997).
[4]J.-D. Polack, "La transmission de l'energie sonore dans les salles" ("Acoustic energy transmission in enclosures"), Ph.D. thesis, Universit du Maine, Le Mans, 1988.
[5]L. Cremer and A. Müller, "Die wissenschaftlichen Grundlagen der Raumakustik" ("Scientific fundamentals of room acoustics"), (S. Hirzel Verlag, Stuttgart, 1978).
[6]P. Rubak and L.-G. Johansen, "Artificial reverberation based on a pseudo-random impulse response II," in *Proceedings of the 106th AES Convention*, Munich (1999).
[7]W. Schmidt and W. Ahnert, "Einfluss der Richtungs-und Zeitdiffusität auf den Raumeindruck" ("Influence of directional and temporal diffusion on the spatial impression of rooms"), Wiss. Z. Techn. Univers. Dresden **22**(2), 313–317 (1973).

[8]M. R. Schroeder, "Natural sounding artificial reverberation," J. Audio. Eng. Soc. **10**, 219–223 (1962).
[9]L. Schreiber, "Was empfinden wir als gleichförmiges Rauschen" ("What do we perceive as uniform noise"), Frequenz **14**(12), 399–403 (1960).
[10]D. Griesinger, "Practical processors and programs for digital reverberation," in *Proceedings of International AES Conference: Audio in Digital Times* (1989).
[11]J.-S. Abel and P. Huang, "A simple, robust measure of reverberation echo density," in *Proceedings of the 121st AES Convention*, San Francisco (2006).
[12]R. Stewart and M. Sandler, "Statistical measures of early reflections of room impulse responses," in *Proceedings of the International Conference on Digital Audio Effects (DAFx-07)*, Bordeaux, France, 2007, pp. 59–62.
[13]T. Hidaka, Y. Yamada, and T. Nakagawa, "A new definition of boundary point between early reflections and late reverberation in room impulse responses," J. Acoust. Soc. Am. **122**, 326–332 (2007).
[14]G. Defrance, L. Daudet, and J.-D. Polack, "Using matching pursuit for estimating mixing time within room impulse responses," Acta Acust. Acust. **95**(6), 1071–1081 (2009).
[15]A. Lindau, L. Kosanke, and S. Weinzierl, "Perceptual evaluation of model- and signal-based predictors of the mixing time in binaural room impulse responses," J. Audio Eng. Soc. **60**(11), 887–898 (2012).
[16]P. Götz, A. Silzle, K. Kowalczyk, and E. Habets, "Diffuseness-based mixing time prediction using a spherical microphone array," in *Proceedings of the International Conference on Spatial Audio (ICSA)*, Erlangen, Germany, 2014.
[17]J. Ahonen and V. Pulkki, "Diffuseness estimation using temporal variation of intensity vectors," *IEEE Workshop Application of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 285–288.
[18]G. Del Galdo, M. Taseska, O. Thiergart, J. Ahonen, and V. Pulkki, "The diffuse sound field in energetic analysis," J. Acoust. Soc. Am. **131**(3), 2141–2151 (2012).
[19]E.-G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography* (Academic Press, London, 1999).
[20]D. P. Jarrett, E. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudo intensity vector," *European Signal Processing Conference (EUSIPCO)*, 2010.
[21]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. Am. **133**, 2711–2721 (2013).
[22]O. Thiergart, G. Del Galdo, and E. Habets, "Signal-to-reverberant ratio estimation based on the complex spatial coherence between omnidirectional microphones," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, pp. 309–312.
[23]J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the sound field," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009, pp. 1781–1784.