PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

3-D sound image localization in reproduction of 22.2 multichannel audio based on room impulse response generation with vector composition

Kaige ZHENG[(1)], Misaki OTSUKA[(2)], Takanobu NISHIURA[(3)]

[(1)]Ritsumeikan University, Japan, is0446ex@ed.ritsumei.ac.jp
[(2)]Ritsumeikan University, Japan, is0206pi@ed.ritsumei.ac.jp
[(3)]Ritsumeikan University, Japan, nishiura@is.ritsumei.ac.jp

Abstract

To simplify content creation processes and improve the realism of 22.2 multichannel system, a method to localize sound images is derived in this paper, which is an improvement of the method based on vector base amplitude panning. The sound image localization can be realized by a specific direction and distance. A sound field is simulated to make spatial impressions in 22.2 multichannel reproduction, which includes reflection and reverberation, and the simulation approach is based on room impulse response generation with vector composition. To control the sound pressure on the ears of the listener, the amplitude of the input signal is attenuated in advance by the distance of sound image. Evaluation experiments were carried out both subjectively and objectively with binaural recording. The improvement in the reproduction of sound image's distance is realized, while the direction of the sound image stays the same as the method based on vector base amplitude panning.

Keywords: 22.2 multichannel, Sound localization

1 INTRODUCTION

With the development of 3-D sound field reproduction technologies, home theater systems are able to create a more realistic experience. In Japan, NHK (Japan Broadcasting Corporation) has developed the 8K Super Hi-Vision standard with 22.2 multichannel audio system [1] in recent years. The traditional way of creating create surround audio contents is using a panning tool to mix the sound signal, but it is difficult to do so in 22.2 multichannel. Besides, the position of a sound image, including distance and direction, is also hard to control with panning. To solve these problems and make the creation process easier, a new method to create contents in 22.2 multichannel sound fields has to be developed.

To keep the compatibility with old surround systems, the object-based rendering method Vector Base Amplitude Panning (VBAP) [2] is used. With the position of sound images, the output signal can be generated automatically based on the configuration of target loudspeaker system. This method focus on the direct sound of objects, therefore the direction information of sound images is easily generated, but the distance can not be controlled precisely.

In this paper, a method to localize sound images based on room impulse response (RIR) generation is derived. This method is described in 3 sections: amplitude correction for input signal, RIR generation, and output signal generation. The input signal's amplitude is adjusted by the distance of the sound image, and the RIR of each channel is generated based on the characteristics of the target sound field and the position of the sound image. This part can reproduce reflections and reverberations that are key to human perception of the distance of the sound source. The output signal is generated by the convolution of RIR and the input signal that has been adjusted in the first part.

2 22.2 MULTICHANNEL AUDIO

The loudspeaker arrangement for a 22.2 multichannel sound system is shown in Fig. 1. The numbers in the figure are the azimuth angle of loudspeakers. With the loudspeakers being placed in elevation directions, a 3-D

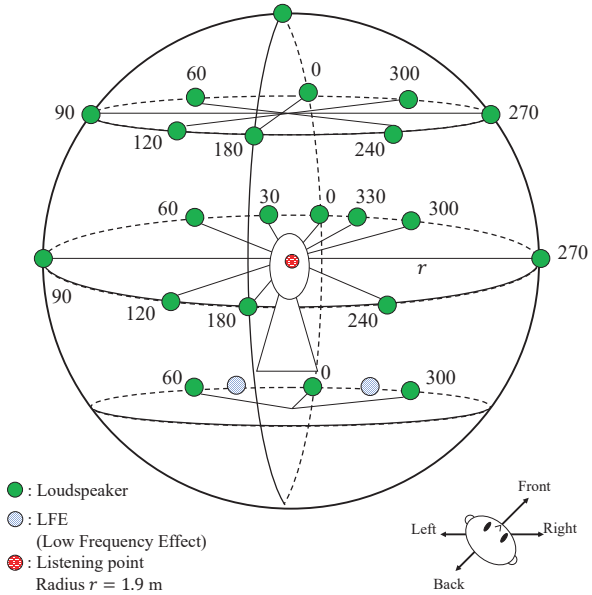

Figure 1. The loudspeaker arrangement for 22.2 multichannel sound system.

sound field can be reproduced by this system, and sound images can be located from all positions.

3 SOUND IMAGE LOCALIZATION BASED ON VECTOR BASE AMPLITUDE PANNING

In the VBAP based method presented in this section, 3 loudspeakers, which are the closest loudspeakers to the position of the sound image, is used to localize the sound image, and the gain factors for these loudspeakers are calculated from the position vector of the sound image.

When the loudspeakers are selected, the gain factors are calculated as

$$\mathbf{p} = (p_x\, p_y\, p_z)^{\mathrm{T}}, \tag{1}$$

$$\mathbf{q}_i = (q_{i,x}\, q_{i,y}\, q_{i,z})^{\mathrm{T}}, \tag{2}$$

$$\mathbf{g} = (g_1\, g_2\, g_3)^{\mathrm{T}}, \tag{3}$$

$$\mathbf{p} = \mathbf{q}_1 g_1 + \mathbf{q}_2 g_2 + \mathbf{q}_3 g_3, \tag{4}$$

$$\begin{pmatrix} p_x \\ p_y \\ p_z \end{pmatrix} = \begin{pmatrix} q_{1,x} & q_{2,x} & q_{3,x} \\ q_{1,y} & q_{2,y} & q_{3,y} \\ q_{1,z} & q_{2,z} & q_{3,z} \end{pmatrix} \begin{pmatrix} g_1 \\ g_2 \\ g_3 \end{pmatrix}, \tag{5}$$

and

$$\begin{pmatrix} g_1 \\ g_2 \\ g_3 \end{pmatrix} = \begin{pmatrix} q_{1,x} & q_{2,x} & q_{3,x} \\ q_{1,y} & q_{2,y} & q_{3,y} \\ q_{1,z} & q_{2,z} & q_{3,z} \end{pmatrix}^{-1} \begin{pmatrix} p_x \\ p_y \\ p_z \end{pmatrix}, \tag{6}$$

where $\mathbf{p}$ is the position vector for the sound image, which points to the center of gravity of the selected loudspeakers, $\mathbf{q}_i$ is the position vector for the $i(\in \{1,2,3\})$th loudspeaker, and $\mathbf{g}$ is the gain factor vector.

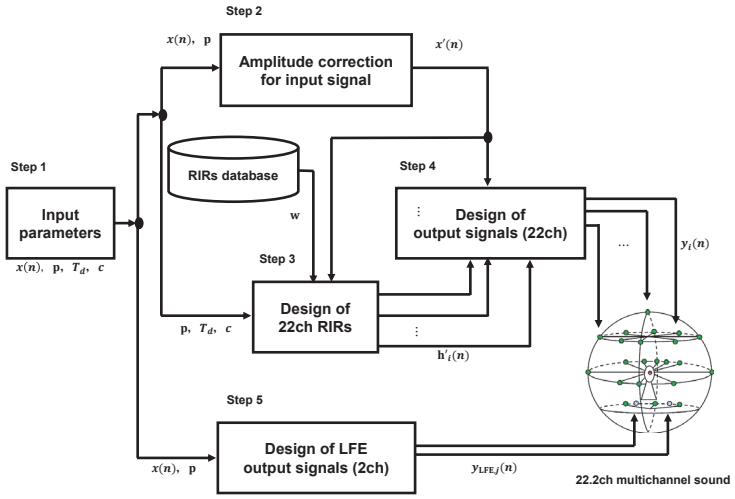

Figure 2. Overview of the proposed method.

The output signal for each channel is generated by the multiplication of the input signal and the gain factors. This method is used to render the direct sound of the sound image, but due to the position of the sound image is limited in the area where the selected loudspeakers are, it is hard for the listener to tell the distance of the sound image. To improve the presence of rendered contents, a new method, which should be able to control the distance of the sound image, needs to be developed.

4 SOUND IMAGE LOCALIZATION BASED ON ROOM IMPULSE RESPONSE GENERATION WITH VECTOR COMPOSITION

The overview of the proposed method is shown in Fig. 2. 3 steps are included in this method: amplitude correction for input signal, design of 22ch RIRs, and output signal generation. Human uses sound pressure and reflected sound to perceive the distance of the sound image, which is controlled by the first and second step of this method separately. Amplitude correction is performed based on the 3-D position vector of the sound image, and 22ch RIRs are designed by an adaptive algorithm, which can simulate the desired sound field for reproduction. To reproduce the ILD (Interaural Level Difference) and the ITD (Interaural Time Difference) that are important to human's perception of direction, the RIRs are also processed with amplitude correction and delay.

4.1 Input parameters

As shown in Fig. 2, the input parameters are: the input signal $x(n)$, the time index n, the position vector of the target sound image $\mathbf{p} = (p_x\, p_y\, p_z)^{\mathrm{T}}$, the reverberation time of the desired sound field T_{d}, and the speed of sound c.

4.2 Amplitude correction for input signal

The amplitude correction for input signal is performed as

$$x'(n) = \alpha x(n), \tag{7}$$

$$\alpha = 10^{\frac{\beta}{20}}, \tag{8}$$

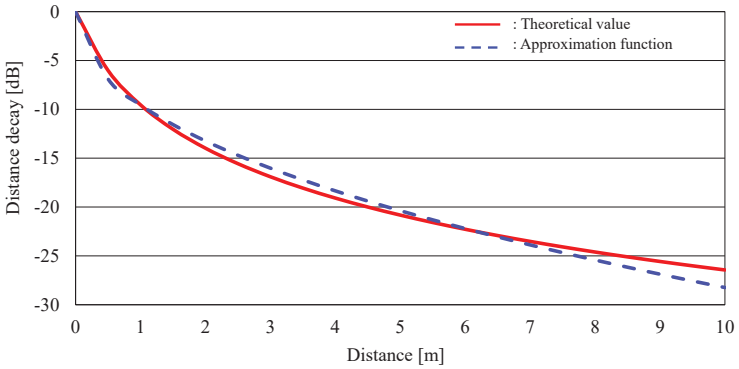

Figure 3. The results of the approximation function for distance attenuation.

and

$$\beta = \begin{cases} -9.55096 \left(\min\limits_{i \in \{1,2,\cdots,22\}} d_i \right)^{0.47082} & \text{if}(d_C < r) \\ 9.55096 \left(\min\limits_{i \in \{1,2,\cdots,22\}} d_i \right)^{0.47082} & \text{otherwise} \end{cases}, \tag{9}$$

where β is the distance attenuation in dB from the sound image to the ith loudspeaker, which is calculated by an approximate function, d_i is the distance between the ith loudspeaker and the sound image, d_C is the distance between the center of the 22.2 multichannel system and the sound image, and r is the radius of the 22.2 multichannel system. The sound pressure of the sound image is controlled by distance by performing this step, and the output signal is $x'(n)$. The distance attenuation from Eq. (9) is compared with theoretical value, and the results are shown in Fig. 3.

4.3 Design of 22ch RIRs

The proposed method generates RIRs by applying subsequent reverberation to early reflection, which is assumed to be uncorrelated. The early reflection, which is equal to estimated RIRs, is generated by an adaptive algorithm, which is written as

$$\mathbf{w} = [w_0, w_1, \cdots, w_{N-1}]^{\mathrm{T}}, \tag{10}$$

$$\mathbf{h}_i(n) = [h_{i,0}(n), h_{i,1}(n), \cdots, h_{i,N-1}(n)]^{\mathrm{T}}, \tag{11}$$

$$\mathbf{x}'(n) = [x'(n), x'(n-1), \cdots, x'(n-N+1)]^{\mathrm{T}}, \tag{12}$$

$$a(n) = \mathbf{w}^{\mathrm{T}} \mathbf{x}'(n), \tag{13}$$

$$a'(n) = \sum_{i=1}^{22} \left(\mathbf{h}_i(n)^{\mathrm{T}} \mathbf{x}'(n) \right), \tag{14}$$

and

$$e(n) = a'(n) - a(n), \tag{15}$$

where $\mathbf{w}$ includes the measured RIRs of the reproduction sound field. The RIRs are selected from an RIR database. N(=80 ms) [3] is the length of $\mathbf{h}_i(n)$, which are the estimated RIRs. $\mathbf{h}_i(n)$ is adapted based on LMS (Least-mean-square) algorithm, where $a'(n)$ is the output signal, $a(n)$ is the target signal, and $e(n)$ is the error at current sample n. The algorithm is repeated based on the correction amount $e(n)$, which is written as

$$\mathbf{h}_i(n+1) = \mathbf{h}_i(n) + 2\mu e(n) g_i' \mathbf{x}'(n), \tag{16}$$

where μ is the step size, and g_i' is the normalized gain factor. The algorithm for calculating the normalized gain factor g_i' is based on vector composition, which is written as

$$\begin{pmatrix} p_x \\ p_y \\ p_z \end{pmatrix} = \begin{pmatrix} l_{1,x} & l_{2,x} & \cdots & l_{22,x} \\ l_{1,y} & l_{2,y} & \cdots & l_{22,y} \\ l_{1,z} & l_{2,z} & \cdots & l_{22,z} \end{pmatrix} \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_{22} \end{pmatrix}, \tag{17}$$

and

$$\mathbf{L} = \begin{pmatrix} l_{1,x} & l_{2,x} & \cdots & l_{22,x} \\ l_{1,y} & l_{2,y} & \cdots & l_{22,y} \\ l_{1,z} & l_{2,z} & \cdots & l_{22,z} \end{pmatrix}. \tag{18}$$

It is assumed that the position vector of the target sound image $\mathbf{p}$ points to the center of gravity of the selected loudspeakers. The gain factor g_i can be calculated by Eq. (17), but the matrix $\mathbf{L}$ is a nonsingular matrix, and therefore has no inverse. In this algorithm, the gain factor g_i is calculated with $\mathbf{L}^+$, which is the pseudo inverse matrix of $\mathbf{L}$. It is written as

$$\begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_{22} \end{pmatrix} = \begin{pmatrix} l_{1,x} & l_{2,x} & \cdots & l_{22,x} \\ l_{1,y} & l_{2,y} & \cdots & l_{22,y} \\ l_{1,z} & l_{2,z} & \cdots & l_{22,z} \end{pmatrix}^+ \begin{pmatrix} p_x \\ p_y \\ p_z \end{pmatrix}, \tag{19}$$

and

$$\mathbf{L}^+ = \begin{pmatrix} l_{1,x} & l_{2,x} & \cdots & l_{22,x} \\ l_{1,y} & l_{2,y} & \cdots & l_{22,y} \\ l_{1,z} & l_{2,z} & \cdots & l_{22,z} \end{pmatrix}^+. \tag{20}$$

The gain factors are inversed in channels that are symmetric with respect to the center of the 22.2 multichannel system. To prevent cancellation in low-frequency band, a normalization process is performed as

$$g_i' = \frac{g_i + |g_{\min}|}{g_{\max} + |g_{\min}|}, \tag{21}$$

where $g_{\max}$ and $g_{\min}$ are the maximum and minimum values of g_i. The output RIRs $\mathbf{h}'_i(n)$ are calculated as

$$\mathbf{h}'_i(n) = \begin{bmatrix} \mathbf{h}_i(n) \\ \mathbf{r}(n) \end{bmatrix}, \tag{22}$$

and

$$\mathbf{r}(n) = [r(n), r(n-1), \cdots, r(n-(T_d - N)+1)]^{\mathrm{T}}, \tag{23}$$

which is the result of applying subsequent reverberation to $\mathbf{h}_i(n)$. The subsequent reverberation is generated from exponentially attenuated white noise. [4]

4.4 Design of output signals

The output signals $y_i(n)$ and $y_{\mathrm{LFE},j}(n)$ are designed based on the RIRs which are calculated in the last section. The output signals for LFE (Low Frequency Effect) channels in the 22.2 multichannel system are generated by an LPF (Low-Pass Filter). The algorithm is written as

$$y_i(n) = \mathbf{h'}_i(n)^{\mathrm{T}}\mathbf{x''}(n), \tag{24}$$

$$y_{\mathrm{LFE},j}(n) = \mathbf{b}^{\mathrm{T}}\mathbf{x''}\left(n - \frac{d_j}{c}\right), \tag{25}$$

$$\mathbf{b} = [b_0, b_1, \cdots, b_{T_d-1}]^{\mathrm{T}}, \tag{26}$$

and

$$\mathbf{x''}(n) = [x'(n), x'(n-1), \cdots, x'(n-T_d+1)]^{\mathrm{T}}, \tag{27}$$

where $\mathbf{x''}(n)$ is a T_d length of signal, which is extracted from the corrected signal $x'(n)$ in Eq. (7). $j(\in\{1,2\})$ is the index for LFE loudspeakers, and $\mathbf{b}$ is the LPF for LFE signals. By reproducing these signals through all the loudspeakers simultaneously, the localization of the sound image is realized, with the direction and the distance well controlled.

5 OBJECTIVE EVALUATION EXPERIMENT

5.1 Experimental conditions

In this experiment, the objective evaluation of the reproduction accuracy of the sound image was carried out by calculating the ILD (Interaural Level Difference) [5]. It shows the difference between the sound pressure of two ears, which are represented by $s_{\mathrm{left}}(n)$ and $s_{\mathrm{right}}(n)$. It is written as

$$\mathrm{ILD} = 20\log 10\frac{|\sum_{n=0}^{N_s-1} s_{\mathrm{left}}(n)|}{|\sum_{n=0}^{N_s-1} s_{\mathrm{right}}(n)|}, \tag{28}$$

where N_s is the length of binaural signals. Based on the ILD (=0 [dB]) when the virtual sound image exists in front of the listener, the ideal tendency is that the ILD rises when the sound image is on the left of the listener and decreases when on the right.

To evaluate the reproduction accuracy of the distance of the sound image, the sound pressure ratio of the binaural signals was calculated as

$$E = 10\log 10\frac{\sum_{n=0}^{N_s-1} s_e^2(n)}{\sum_{n=0}^{N_s-1} s_e'^2(n)}, \tag{29}$$

where $s'_e(n)$ is the binaural signal $e(\in\{\mathrm{left},\mathrm{right}\})$ when the sound image is localized at the reference point. The sound pressure ratio E tends to decrease because the amount of distance attenuation increases as the distance of the sound image becomes farther than the reference point.

In this experiment, the conventional VBAP method, the proposed method, and the real environment (recorded as binaural sound using Neumann KU100 microphone) were compared. To avoid the influence of the physical loudspeakers, the signals are not recorded (except the real environment), but generated as

$$s_e(n) = \sum_{i=1}^{22} \mathbf{k}_{i,e}{}^{\mathbf{T}}(n)\mathbf{y}_i(n) + \sum_{j=1}^{2} \mathbf{k}_{j,e}{}^{\mathbf{T}}(n)\mathbf{y}_{\mathrm{LFE},j}(n), \tag{30}$$

where $\mathbf{k}_{i.e}(n)$ is the BRIR (Binairal Room Impulse Response) from the ith loudspeaker to the binaural microphone, which was measured in advance. Other conditions for this experiment is shown in Table 1.

The direction is in counterclockwise with the front of the listener being 0. The distance represents the distance from the center of the 22.2 multichannel system. The input signal was a white noise in 16 [bit] / 48 [kHz]. The step size μ of the adaptive algorithm in the proposed method was 10^{-12}. The adaptive algorithm was repeated 400 times.

Table 1. Experimental conditions for objective evaluation

Environment	Office room
Reverberation time	$T_{[60]}$ = 650 [ms]
Direction	0, 30, ···, 180 [deg.]
Distance	1.0, 2.0, ···, 5.0 [m]
Height	1.4 [m]

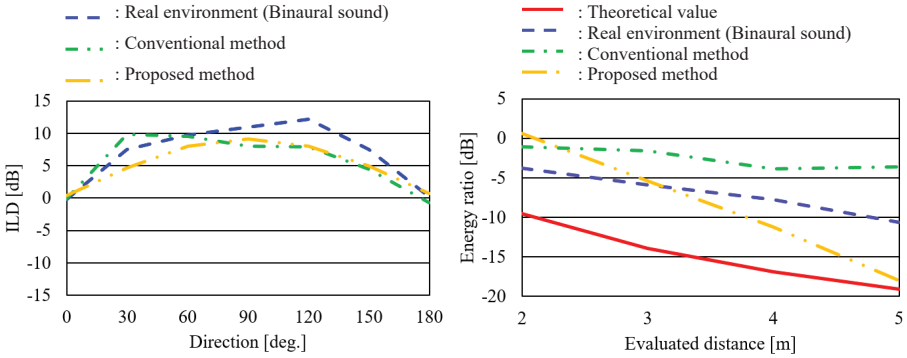

Figure 4. Left: Experimental results by direction with a distance of 1.0 [m]. Right: Experimental results by distance with a direction of 30 [deg].

5.2 Results

The results are shown in Fig. 4. When the distance of the sound image was fixed as 1.0 [m], both the conventional method and the proposed method can reproduce the same transition of ILD that approximates the real environment. No significant difference was identified between the real environment and the two methods with a t-test performed, where the significance level is 5%. When the direction of the sound image was fixed as 30 [deg], the distance of the reference point in Eq. (29) was fixed as 1.0 [m], and the theoretical value was calculated using Eq. (9), the energy ration tends to decrease as the distance of the sound image going farther, but due to the fact that the distance controlling is insufficient in the conventional method, the results of the conventional method are relatively flat, compared to the proposed method, which leads to the conclusion that the proposed method is able to control the distance of the sound image. Comparing to the real environment, the energy ratio is greatly reduced with the proposed method, showing that the amplitude correction in the proposed method needs to be optimized for the reproduction environment.

6 SUBJECTIVE EVALUATION EXPERIMENT

6.1 Experimental conditions

In this experiment, the subjective evaluation of the reproduction accuracy of the sound image was carried out, which focused on the distance of the sound image. When the direction of the sound image was fixed as 0 [deg], 5 subjects were asked to tell the distance of the sound image. The subjects were also asked to listen to evaluation signals of 1.0 [m] and 5.0 [m] before the experiment began. Sony MDR-CD900ST headphone was used for binaural reproduction. Other conditions for this experiment are shown in Table 2.

6.2 Results

The results are shown in Table 3, which represent the average correct answer rate of the 5 subjects. The improvement on the reproduction of the distance of the sound image was obvious when the distance was set

Table 2. Experimental conditions for subjective evaluation

Condition	Binaural sound	Conventional and proposed method
Environment	Sound proof room	22.2ch multichannel reproduction room
Temperature	18.7 [°C]	20.3 [°C]
Humidity	30.1 [%]	26.1 [%]
Ambient noise level	$L_A = 30.4$ [dB]	$L_A = 38.0$ [dB]

Table 3. Experimental results for subjective evaluation

Presented distance [m]	Binaural sound [%]	Conventional method [%]	Proposed method [%]
1	60	54	50
3	34	28	28
5	32	2	65

to 5 [m]. Combining the results from the objective evaluation, it is clear that the proposed method has better control of the distance of the sound image in the range of 3 to 5 [m].

7 CONCLUSIONS

A new method to localize sound images based on RIR generation is proposed in this paper, and the improvement of this method is confirmed comparing the conventional method, which is based on VBAP. This new method can be used to simplify the content creation process for 22.2 multichannel systems. In the future, the accuracy of sound image localization based this method can be improved by optimizing parameters of the adaptive algorithm for other reproduction environments.

ACKNOWLEDGEMENTS

This research was partly supported by Ritsumeikan Global Innovation Research Organization, JST COI and JSPS KAKENHI Grant Numbers JP18K19829 and JP19H04142.

REFERENCES

[1] K. Hamasaki, T. Nishiguchi, R. Okumura, Y. Nakayama and A. Ando, "A 22.2 Multichannel Sound System for Ultra-High-Definition TV (UHDTV)," SMPTE Motion Imaging Journal, vol. 117, no. 3, pp. 40-49, 2008.

[2] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," Journal of the Audio Engineering Society, vol. 45, no. 6, pp. 456-466, 1997.

[3] B. G. Shinn-Cunningham, N. Kopco and T. J. Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," The Journal of the Acoustical Society of America, vol. 117, no. 5, pp. 3100-3115, 2005.

[4] T. Yoshimura, Y. Wakabayashi, T. Fukumori, M. Nakayama and T. Nishiura, "Sound localization with distance perception based on simulated room impulse response using 16ch head-enclosed loudspeaker-array," in RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing 2017 (NCSP'17), pp. 421-424, 2017.

[5] N.Roman, D. Wang and G. J. Brown, "Speech segregation based on sound localization," Journal of the Acoustical Society of America, vol. 114, no. 4, pp. 2236–2252, 2003.