METHOD OF ESTIMATING DIRECTION OF ARRIVAL OF SOUND SOURCE FOR MONAURAL HEARING BASED ON TEMPORAL MODULATION PERCEPTION

Nguyen Khanh Bui, Daisuke Morikawa and Masashi Unoki

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, Japan

ABSTRACT

Although humans are capable of using monaural and modulation cues for sound localization, it is not yet clear how they can use that information to estimate the direction of arrival (DOA) of a sound source in 3D space. Our previous study revealed that the head-related modulation transfer function (HR-MTF) contains significant trends and features, which can be used for DOA estimation. This paper proposes a method of estimating the DOA in a 3D space by using the monaural modulation spectrum (MMS), based on the concept of modulation transfer function (MTF) and auditory perception of temporal modulation. We carried out over $51,840$ simulations with several signal types and multiple subjects to simultaneously estimate the azimuth and the elevation of an incoming sound source. The root mean square error (RMSE) was derived to evaluate the accuracy of monaural DOA estimates. Our results indicated that the proposed method could adequately estimate the DOA in 3D space with an overall mean RMSE of 21.9 degrees.

Index Terms— Direction of arrival, temporal modulation, auditory perception, monaural modulation spectrum, head-related modulation transfer function

1. INTRODUCTION

The human ability of identifying the source of sounds arriving at the ear has been investigated in the field of acoustic signal processing for many years. Although most studies on direction of arrival (DOA) have adopted binaural cues, it has recently been reported that we are able to use monaural cues for DOA in absence of binaural cues [1]. Therefore, estimating monaural DOA based on human hearing mechanism has been considered to be challenging problem and a solution to it may prove to be beneficial.

It has also been suggested that modulation cues play an important role in monaural DOA estimates [2]. Based on this, some studies have indicated interest in using the monaural modulation spectrum (MMS) in estimating DOA. Kliper et al. proposed a method that used machine learning to directly identify DOA from MMS [3]. However, as it was not clear how MMS could be used by humans, it could not explain our hearing mechanism. Ando et al. reported that MMS shapes were approximately drawn as arcs with azimuth variations [4, 5] by using the modulation transfer function (MTF). They also demonstrated the feasibility of using MMS to estimate DOA on the horizontal plane. Nevertheless, their study did not rigorously investigate the head-related modulation transfer function (HR-MTF), which contains important DOA-related characteristics.

This work was supported by a Grant in Aid for Scientific Research, Innovative Areas (No. 16H01669) from MEXT, Japan and the Kayamori Foundation. The authors would like to thank Prof. Tatsuya Hirahara (Toyama Prefectural University) and RIEC of Tohoku University for the use of BEM and RIEC HRTF dataset.

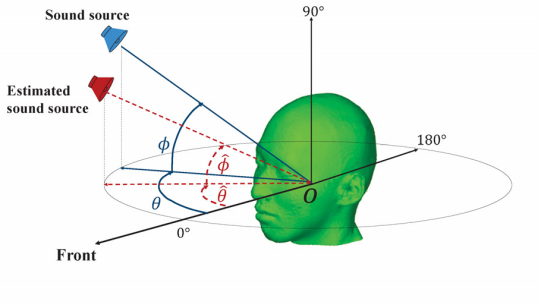

Figure 1: DOA estimation conventions in 3D space.

Our previous study explained the transfer function with different datasets on both horizontal and median planes to enable the HR-MTF to be understood [6]. The results indicated that HR-MTF has significant trends and features that could be useful for monaural DOA estimation. The curves on HR-MTF could clearly be seen from 100 to 350 degrees of azimuth, and from -30 to 100 and 120 to 220 degrees of elevation. We also found that there were some notable features such as a large peak from 80 to 110 degrees of azimuth and dips at 210 degrees of azimuth and 110 degrees of elevation. The same tendency could similarly be found with regard to modulation frequencies. Thus, this was the basis for further investigations into the temporal modulation cues for use on monaural DOA estimates.

Our main aim in this study was to propose a method of estimating monaural DOA in 3D space, based on the human hearing mechanism. Section II introduces the concepts of modulation perception with regard to the MTF concept. Section III describes our estimation strategy. We confirmed the tendency of HR-MTF with consideration of temporal modulation perception, and then investigated the MMS of speech to find important features to be used for DOA estimation. Section IV explains our proposed method, which we evaluated and discuss in Section V. Finally, the paper is summarized in Section VI.

2. CONCEPTS OF MODULATION PERCEPTION

It has reported that we have modulation frequency selectivity in the auditory system and the temporal amplitude envelope is processed by a modulation filterbank [7, 8]. Therefore, the amplitude modulation domain is a very significant dimension in hearing. Recent studies by Greenburg, Atlas and Hermansky have found that the modulation spectrum conveys linguistic information on speech [9]. More-

over, it seems that modulations at low frequencies play significant roles on carrying the information of speech [10, 11].

On the other hand, MTF was introduced to measure the effect of enclosure on speech intelligibility [12] since it is necessary to evaluate speech transmission. In this case, we applied this concept to our approach. Let us assume that we have an arbitrary source signal arriving from a distant source location to one of our ears as shown in Fig. 1. Before it reaches an eardrum, the signal is reflected and diffracted by the human body structure. These changes are captured by the head-related impulse response (HRIR), which contains spatial information.

Based on the MTF concept, let $x(t)$, $h(t,\theta,\phi)$, and $y(t,\theta,\phi)$ correspond to the sound source signal, HRIR, and observed signal. The θ and ϕ are the azimuth and elevation in these variables, which describe the DOA of the approaching sound as shown in Fig. 1. We can simply take the convolution of the source signal $x(t)$ and HRIR $h(t,\theta,\phi)$ according to this concept, to ascertain the sound pressure measured at the ear drum position, which is $y(t,\theta,\phi)$.

The temporal power envelope of the observed signal, $e_y^2(t,\theta,\phi)$, in the modulation domain, on the other hand, can be calculated as:

$$e_y^2(t,\theta,\phi) = e_h^2(t,\theta,\phi) * e_x^2(t), \tag{1}$$

where $e_x^2(t)$ and $e_h^2(t,\theta,\phi)$ are the power envelopes of $x(t)$ and the HRIR $h(t,\theta,\phi)$ [13]. Eq. (1) can be represented in the modulation-frequency domain as:

$$E_y(f_m,\theta,\phi) = E_h(f_m,\theta,\phi)E_x(f_m), \tag{2}$$

where $E_x(f_m)$, $E_h(f_m,\theta,\phi)$, $E_y(f_m,\theta,\phi)$, and f_m are the MMS of $x(t)$, HR-MTF of the HRIR $h(t,\theta,\phi)$, the MMS of $y(t,\theta,\phi)$, and the modulation frequency. From the power envelopes of $h(t,\theta,\phi)$, HR-MTF is calculated by:

$$E_h(f_m,\theta,\phi) = \int_0^\infty e_h^2(t,\theta,\phi)\exp(-j2\pi f_m t)dt. \tag{3}$$

The $e_y^2(t,\theta,\phi)$ in this method of estimation was extracted by:

$$e_y^2(t,\theta,\phi) = \mathrm{LPF}\Big[|y(t,\theta,\phi) + j\mathrm{Hilbert}[y(t,\theta,\phi)]|^2\Big], \tag{4}$$

where Hilbert[·] is the Hilbert transform and LPF[·] is low-pass filtering. This equation is based on the calculation of instantaneous amplitude, and low-pass filtering is used to remove the higher modulation-frequency components in the power envelope as post-processing. Finally, $e_y^2(t,\theta,\phi)$ is transformed to $E_y(f_m,\theta,\phi)$ using fast Fourier transform (FFT).

3. ESTIMATION STRATEGY

Human being can only obtain and process the observed signal in general sound localization. However, we can perceive important monaural/binaural cues of DOA information from this particular input. Therefore, monaural DOA estimation would also only account for the observed signal based on the human hearing mechanism.

The monaural modulation spectrum of the observed signal, $E_y(f_m,\theta,\phi)$, can be obtained from Eq. (2) from the multiplication of the HR-MTF, $E_h(f_m,\theta,\phi)$, and the MMS of the source signal, $E_x(f_m)$. We know that the HR-MTF contains the DOA information in the MMS of the observed signal; thus, our main interest was to extract this useful information from the observed signal's MMS. However, the MMS of the source signal, which is also a component in Eq. (2), certainly affects the observed signal's MMS.

Our strategy to estimate the monaural DOA according to these considerations was comprised of three steps:

First, we confirmed our previous study [6] results on HR-MTF tendencies in a range of meaningful modulation frequencies for human hearing. This took into consideration auditory perception of temporal modulation to understand what modulation frequencies are useful for monaural DOA.

Second, we studied the MMS of the source signal. Each type of source signal should have different characteristics that affect the observed signal. In this case, we used speech, which is a natural signal that humans use for DOA. The features that we found from the speech MMS could then be applied to create similar artificial signals for the next step.

Finally, we propose a method of estimating the monaural DOA in 3D space based on this preliminary knowledge. We expected to obtain useful spatial information from the observed signal by using the features of MMS of the source signal.

3.1. HR-MTF tendencies

It was obvious from our previous study to see that the HR-MTF characterized important tendencies that could be used for monaural DOA by conducting analyses with different datasets on both elevation and median planes [6]. The HR-MTF was approximately drawn as curves in three ranges. The first range was from 110 to 350 degrees of azimuth, while the others were from -30 to 100 degrees and 120 to 220 degrees of elevation. Other ranges could also be considered to be in the curves, however, since they were additionally affected by different passive effects, they were more difficult to observe. The results also shared a consideration on the relationship between HR-MTF and the range of meaningful modulation frequency via geometrical analysis. The trends and features of HR-MTF seemed to remain identical with both the azimuth and elevation with different modulation frequency variations.

We know that at higher frequencies, smaller parts of the head will have more of an effect on the HR-MTF based on the characteristics of sound moving in space. In addition, modulation frequency should be related to amplitude modulation and carriers. This suggests that a sufficiently low modulation frequency would range from 1 Hz to 60 Hz. Moreover, it is known that the fluctuation strength is largest for 4 Hz in consideration with the perception of temporal amplitude modulations and reaches zero at 32 Hz [14]. Also, the fluctuation strength depends very little on the frequency of sinusoidal carriers [15]. Therefore, a meaningful scope for modulation frequency was appropriately chosen in this study to range from 1 Hz to 40 Hz. This consideration will be the basis for building the method of monaural DOA estimation.

Figure 2 plots the HR-MTF with a range of lower modulation frequencies. The abscissas indicate the azimuth or elevation, while the ordinates depict the modulation depth in decibels (dB). The three curves inside plot the HR-MTF with different frequency variations. As expected, the similar tendencies that we found could also be seen in the human-perception-based modulation frequencies of 4, 16, and 32 Hz.

3.2. Speech characteristics on modulation spectrum

A corpus of speech called "Familiarity-controlled word lists 2007" has been used [16] to investigate speech MMS characteristics. This corpus was developed to assess hearing abilities under daily-life conditions and was recorded in Japanese. A total of 6.400 signals have been analyzed with the process.

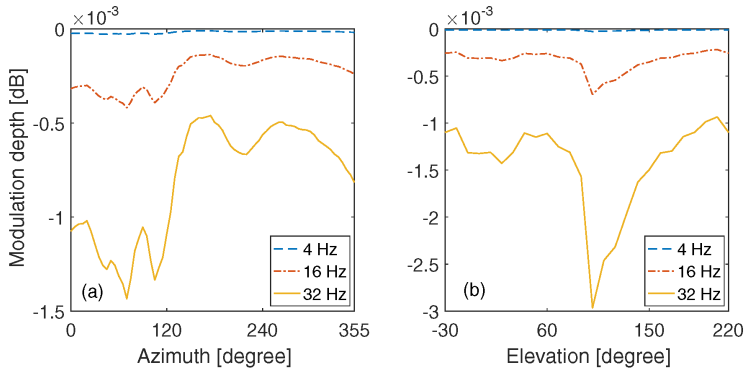

Figure 2: HR-MTF tendencies with different modulation frequencies.

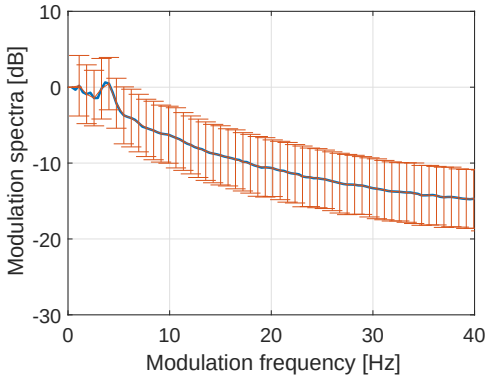

Figure 3: Mean and standard deviation of speech modulation spectra.

Figure 3 shows the averaged modulation spectra of speech obtained from the FW07 corpus, where the horizontal axis indicates the modulation frequencies, while the vertical axis indicates the modulation spectra. The red solid bars are the standard deviations in the dataset, while the solid blue curve plots the mean values.

It can be seen that there are two characteristics that are similar in the corpus. The first characteristic is that at the modulation frequency of 4 Hz, there is a dominant peak, or the largest peak across the modulation frequency range. This might have resulted from the duration of a temporal unit called "mora" in Japanese speech [17], where the typical mora duration is from 0.2 to 0.25 s. Therefore, it is obvious to see the dominant change in the modulation spectra is at the modulation frequency of 4 Hz.

The second characteristic is the spectral tilt of the modulation spectra from 10 to 40 Hz. After reaching the highest peak at 4 Hz, the MMS decreases as the modulation frequency increases. This could be why the fluctuation strength for human perception is largest for 4 Hz and reaches zero at 32 Hz. We conducted regression analysis (by using polynomial fitting from 10 to 40 Hz) to measure the spectral tilt of the MMS in this frequency range. The tilt per octave was -4.69 (dB/oct).

4. PROPOSED METHOD

The proposed method for estimating monaural DOA is outlined in the block diagrams in Fig. 4. There were two phases that were essential in this study.

The power envelope of the observed signal, $e_y^2(t,\theta,\phi)$, was calculated in the first phase with Eq. (4) where an LPF with a cut-off frequency of 50 Hz was used. Then, the envelope, $e_y^2(t,\theta,\phi)$, was transformed to the modulation spectra of observed signal $E_y(f_m,\theta,\phi)$ using FFT. However, the auto-correlation function (ACF) was used to determine the dominant modulation frequency, f_{dm} in $E_y(f_m,\theta,\phi)$, to obtain the peak feature, $P_y(\theta,\phi)$. Moreover, we obtained the slope feature, $S_y(\theta,\phi)$, by regression analysis.

Polynomial regression was then used to model the peak and slope features into $G_p(\theta,\phi)$ functions for the former and $G_s(\theta,\phi)$ functions for the latter. Figures 5 and 6 show these features that were obtained from simulations, along with the fitting results. It can be seen that the features could be approximately fitted using polynomial regression. This suggests that the curves, or the important tendencies of the HR-MTF, were successfully transferred into the MMS of the observed signal, which is useful for DOA estimation.

The peak and slope of MMS of the observed signal was obtained in the estimation phase with respect to the first phase in the same manner. Then, the DOA is calculated by:

$$(\hat{\theta},\hat{\phi}) = \underset{\Delta\theta\geq 0,\Delta\phi\geq 0}{\arg\min} \left(\sqrt{(G_p(\theta,\phi)-P_y)^2+(G_s(\theta,\phi)-S_y)^2}\right), \tag{5}$$

where $\hat{\theta}$ is the estimated azimuth and $\hat{\phi}$ is the estimated elevation.

Information of head movement was added in this study to determine the final estimation candidate in Eq. (5). Figures 5 and 6 show that the MMS values increase or decrease according to the azimuth or elevation, which can be regarded as the movement and shape of the head. It seems that when the head is moving clockwise (in which the azimuth is increasing), particularly in Fig. 5, the peak value would likely decrease in the azimuth range from 0 to 90 degrees and from 290 to 355 degrees. However, it would decrease with the azimuth from 90 to 290 degrees. Moreover, it can be seen from Fig. 6 that when the head is turned up (along with the elevation increasing), the slope value decreases in the elevation range from -30 to 15 degrees. After that, when the elevation is more than 15 degrees, the slope value increases up to 80 degrees.

The estimated DOA before ($\hat{\theta}$, $\hat{\phi}$) and after ($\hat{\theta}$ + Δ θ, $\hat{\phi}$ + Δ ϕ) would be substituted into the derivatives of peak and slope functions by bearing this consideration in mind. This resulted in an increase or decrease.

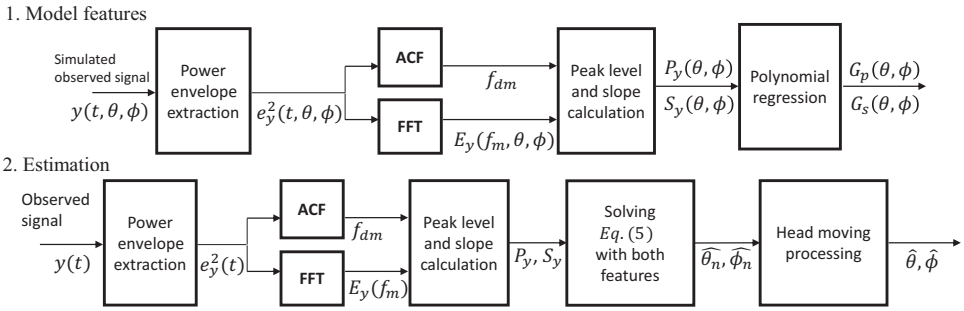

Figure 4: Block diagram of proposed method.

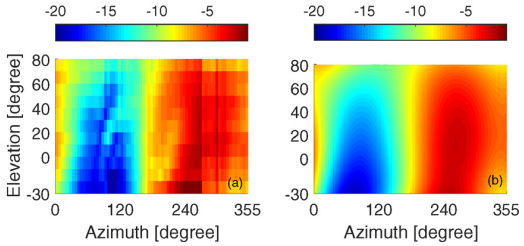

Figure 5: Peak feature $P_y(\theta,\phi)$ (left) and its fitting function $G_p(\theta,\phi)$ (right).

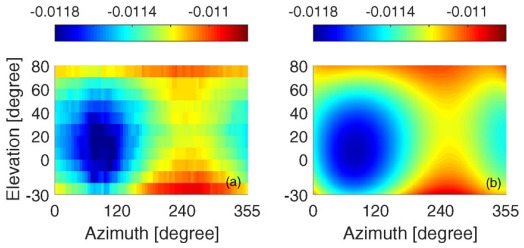

Figure 6: Slope feature $S_y(\theta,\phi)$ (left) and its fitting function $G_s(\theta,\phi)$ (right).

5. EVALUATION

We investigated how the proposed method could be used for monaural DOA estimation in 3D space based on this discussion. HRIRs, $h(t,\theta,\phi)$s, were used from the Research Institute of Electrical Communication head-related transfer functions (RIEC HRTF) dataset recorded by RIEC, Tohoku University [18, 19]. The dataset included 104 subjects (208 ears) at 864 positions. The source distance, r, was 1 m, and the sampling frequency was 48 kHz.

Amplitude modulated (AM) noise was used in our simulations as the source signal $x(t)$ in the test. The AM noise contained a white noise carrier and several modulating signals with a modulation frequency of 4 Hz and a range from 10 to 40 Hz, in which the signals had an artificial slope of -4.69 dB/oct. The length of the signal was also $1,000$ ms. In total, we used $51,840$ artificial signals ($= 10 \times 6$ left ear HRIRs $\times 864$ possible positions in 3D space).

We derive the Root Mean Square Error (RMSE) for each of the test positions to evaluate the proposed method by:

$$\mathrm{RMSE}(\hat{\theta},\hat{\phi}) = \sqrt{\frac{1}{N}\sum((\hat{\theta}-\theta)^2 + (\hat{\phi}-\phi)^2)}, \qquad (6)$$

Table 1: Estimation results in RMSE with 40 azimuth and 20 elevation increments.

Azimuth \ Elevation	-30	-10	10	30	50	70	80
0	11.1	4.9	13.4	10.5	11.2	16.2	11.6
40	13.8	7.5	11.2	15.1	10.5	10.8	6.1
80	29.0	32.5	8.6	27.5	11.3	12.9	6.4
120	15.3	17.1	8.1	16.5	6.9	8.5	7.6
160	12.6	16.4	17.6	17.5	26.1	13.4	14.4
200	31.7	20.9	25.8	16.8	12.9	14.6	18.2
240	18.6	28.5	16.7	23.5	15.2	8.2	9.4
280	35.1	4.5	28.5	14.0	21.6	17.3	16.0
320	7.4	11.5	10.9	13.2	31.5	33.2	30.1
355	12.8	9.8	18.2	18.8	17.3	27.8	32.5
Mean RMSE							**21.9**

where N is the number of simulations in each position (θ,ϕ). Table 1 lists the RMSE results for a larger increment in azimuth and elevation, which were 40 degrees for the former and 20 degrees for the latter. It can be seen that there are no obvious regions that have noticeably better or worse DOA estimates. However, the accuracy of the method does seem to increase near the horizontal plane. The overall mean RMSE of the method was calculated as 21.9 degrees. Since our test dataset (RIEC HRFT) has 5 degrees for azimuth and 10 degrees for elevation resolutions, the method is believed to make better results if the dataset has smaller degree increments.

6. CONCLUSION

We proposed a method for estimating monaural DOA in 3D space in this paper that was based on human perception of temporal modulation. We employed the trends and features previously found on HR-MTF by using the MTF concept as important cues that contained spatial information. We also analyzed speech, which is a natural signal that humans use for DOA, to find important features that could be utilized for the method. We finally evaluated the method by deriving the RMSE for 864 possible positions in 3D space with different HRIRs and signals. The results revealed that the proposed method could adequately estimate the DOA with an overall RMSE of 21.9 degrees. We will plan to improve the proposed method with regard to the accuracy of estimates and robustness against different HRIR datasets and types of the source signal. We will also plan to investigate whether the proposed method can account for the experimental data of monaural DOA by humans.

7. REFERENCES

[1] K. Strelnikov, M. Rosito, and P. Barone, "Effect of Audiovisual Training on Monaural Spatial Hearing in Horizontal Plane," *PloS one*, vol. 6, no. 3, pp. 1-9, 2011.

[2] E. R. Thompson and T. Dau, "Binaural processing of modulation interaural level difference," *J. Acoust. Soc. Am.*, vol.123, no. 2, pp. 1017-1029, 2008.

[3] R. Kliper, H. Kayser, D. Weinshall, I. Nelken, and J. Anemuller, "Monaural azimuth localization using spectral dynamics of speech," *Proc. Interspeech*, pp. 33-36, 2011.

[4] M. Unoki, "Speech Signal Processing Based on the Concept of Modulation Transfer Function -Basis of Power Envelope Inverse Filtering and Its Applications," *J. Signal Processing*, vol. 12, no. 5, pp. 339-348, 2008.

[5] D. Morikawa, M. Ando, and M. Unoki, "Feasibility of Estimating Direction of Arrival Based on Monaural Modulation Spectrum," *Proc. 2015 International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP2015)*, pp. 384-387, Adelaide, Australia, 2015.

[6] N. K. Bui, D. Morikawa, and M. Unoki, "Investigation on the head-related modulation transfer function for monaural DOA," *Advances in Intelligent Information Hiding and Multimedia Signal Processing*, pp. 191-198, 2016.

[7] T. Dau, "Modeling auditory processing of amplitude modulation," PhD thesis, Universität Oldenburg, 1996.

[8] T. Dau, and B. Kollmeier, *Modeling auditory processing of amplitude modulation. I. Detection and masking with narrowband carriers*. J. Acoust. Soc. Am., vol. 102, pp. 2892-2905, 1997.

[9] L. Atlas, S. Greenberg, and H. Hermansky, "The Modulation Spectrum and Its Application to Speech Science and Technology," *Proc. Interspeech2007*, Tutorial, 2007.

[10] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Syllable intelligibility for temporally filtered LPC cepstral trajectories," *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2783-2791, 1999.

[11] R. Drullman, "Temporal envelope and fine structure cues for speech intelligibility," *J. Acoust. Soc. Am.*, vol. 97, no. 1, pp. 585-592, 1995.

[12] T. Houtgast, and H. J. M. Steeneken, "The Modulation Transfer Function in Room Acoustics as a Predictor of Speech Intelligibility," *Acustica*, vol. 28, pp. 66-73, 1973.

[13] M. Unoki, M. Furukawa, K. Sakata, and M. Akagi, "A method based on the MTF concept for dereverberating the power envelope from the reverberant signal," Proc. ICASSP2013, vol. 1, pp. 840-843, 2013.

[14] E. Zwicker, and H. Fastl, *Psychoacoustics*. Berlin: Springer, 1999.

[15] S. Sheft and W. Yost, "Temporal integration in amplitude modulation detection," *J. Acoust. Soc. Am.*, vol. 88, pp. 796-805, 1990.

[16] T. Kondo, S. Amano, S. Sakamoto, Y. Suzuki, "Development of familiarity-controlled word-lists (FW07)," *EICE Tech. Rep.*, vol. 107, pp. 43-48, 2008.

[17] H. Kubozono, "The mora and syllable structure in Japanese: Evidence from speech errors," *Language and Speech*, vol. 32, pp. 249-278, 1989.

[18] The RIEC HRTF Dataset, http://www.riec.tohoku.ac.jp/pub/hrtf/index.html.

[19] K. Watanabe, Y. Iwaya, Y. Suzuki, S. Takane, and S. Sato, "Dataset of Head-related Transfer Functions Measured with a Circular Loudspeaker Array," *Acoustical Science and Technology*, vol. 35, no. 3, pp. 159-165, 2014.