Microphone Position Realignment by Extrapolation of Virtual Microphone

Ryoga Jinzai[1], *Kouei Yamaoka*[1], *Mitsuo Matsumoto*[1], *Takeshi Yamada*[1], *Shoji Makino*[1]
[1]University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8577 Japan
E-mail: s1820656@s.tsukuba.ac.jp, yamaoka@mmlab.cs.tsukuba.ac.jp, makilab-research@tara.tsukuba.ac.jp, takeshi@cs.tsukuba.ac.jp, maki@tara.tsukuba.ac.jp

***Abstract*— In this paper, microphone realignment by phase extrapolation using the virtual microphone technique for reproducing binaural signals with adequate the interaural time differences (ITDs) for a listener is proposed. For a sound source in the horizontal plane, ITDs are a major cues for localizing a sound image. Since ITDs are not considered for headphones listening in conventional amplitude panning in multichannel recording, sound images are localized inside the head (lateralization). A microphone array is applicable to recording signals with time differences corresponding to the directions of sound sources. Since microphones in such an array are closely positioned, the time differences are inappropriate as ITDs for localizing sound images for the sources. In this paper, phase extrapolation using the virtual microphone technique is applied to the virtual realignment of a microphone in such an array for restoring ITD. In the experiments with two speeches as sound sources located at the leftmost and the rightmost positions from the viewpoint of two real microphones positioned 2.83 cm apart. Furthermore, the phase of a signal of a virtual realigned microphone is extrapolated eight times as much as the phase between the two real microphones. Time differences between signals of one of the real microphones and the realigned one are observed to be -500 μs for the source on the left and 500 μs for the source on the right. Furthermore, the interaural cross correlations of the two signals suggest that sound images will be perceived on both the left and right of a listener. In this method, it is expected that prior information on the number of sources and the direction of arrival is not required, and the adjustment of individual differences is easy.**

I. INTRODUCTION

Sound image localization is an auditory event in spatial impression. For sound sources in the horizontal plane, the interaural time difference (ITD or phase difference) and interaural level difference (ILD) are localization cues based on binaural differences [1-3]. ITD is maximum when the sound sources are at the leftmost or rightmost position (approximately 630 μs). A dummy head is available for recording binaural signals with adequate ITD. However, ITD depends on the listener, and a dummy head that is an exact replica of the listener is unrealistic. The individualization for ITD has been studied in [4]. The head-related transfer function (HRTF) is defined as a transfer function between a sound source and the entrance of the ear canal. As HRTF contains all cues for perception on spatial impression, ITD can be extracted from HRTF to interpolate the individualization as reported in [5]. Moreover, the head and torso simulator has been proposed in [6] for preparing an evaluation criterion using the difference between the measured HRTF and the simulated one, which means that ITD can still be regarded as a research topic.

Recently, audio signal processing in the time–frequency domain has been actively proposed. Sound source (musical instrument) separation and re-panning (reallocation) in conventional 2-ch stereophony or multichannel contents have been investigated in [7-10]. Since sound sources (musical instruments) are independently recorded by multichannel microphones and sound images are located by amplitude panning, there is no relationship between the time difference of the microphone signals and the locations of sound sources. Therefore, ITDs between the signals after re-panning cannot be restored in the headphones listening. If the signals are presented to a listener, the listener will perceive sound images inside the head (lateralization).

To solve this issue, if a microphone array (at least a 2-ch array) is applicable, the direction of each sound source from the viewpoint of the array is recorded as the time difference between microphone signals. Those time differences depend on the locations of the sound sources and the distance between the microphones. As the microphones used for the microphone array are adjacently and closely positioned to prevent spatial aliasing, it is obvious that the distance between the microphones is different from the distance between the two ears of a listener. This means that the time difference detected by the microphone array is different from the ITD that the listener perceives. Therefore, if the sound recorded by the microphone array is presented to the listener, the location of the sound source is wrongly perceived. Moreover, an unwrapped phase shift with the virtual microphone technique is applied to localize a sound image laterally in [11]. It is impossible for "phase unwrapping" to move multiple sound images in different directions.

In this study, phase extrapolation by the virtual microphone technique is employed to resolve the misalignment caused by the microphone array [11] and to make the distance between microphones equivalent to the distance between the two ears of a listener to restore ITD. This is confirmed by observing the waveforms of the sound sources and/or interaural cross correlations.

In this method, prior information on the number of sound sources and arrival direction is unnecessary for microphone realignment. Moreover, it can be expected that the individual

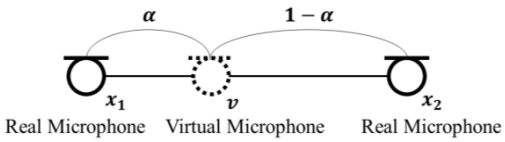

Fig. 1 Arrangement of real and virtual microphones in interpolation technique

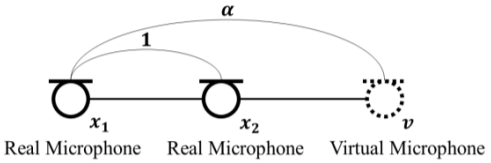

Fig. 2 Arrangement of real and virtual microphones in extrapolation technique

difference in the ITD can be easily adjusted by the extrapolation coefficient α.

II. INTERPOLATION OF VIRTUAL MICROPHONE SIGNAL

In this section, we introduce the interpolation with virtual microphone technique [12] [13].In this technique, a virtual microphone signal $v(\omega, t)$ is generated from the observed signals of two real microphones x_i in the time–frequency domain, where $x_i(\omega, \mathrm{t})$ is the ith microphone signal ($i = 1,2$) at the angular frequency ω in the t th time frame. α is a coefficient of the interpolation of the virtual microphone. The arrangement of real and virtual microphones is shown in Fig. 1.

In an environment where there are multiple sounds arriving from different directions, the relationship between microphone position and waveform seems to be complicated. In this method, by assuming the W-disjoint orthogonality (W-DO) of the observed signals, we simplify the modeling of the relationship. W-DO means the strong sparsity of the signal in the time–frequency domain, assuming that the component from a sound source is dominating one time–frequency slot of discrete short time Fourier transform (STFT). By assuming W-DO, when multiple sounds arrive, they can be regarded as a single sound in each time–frequency bin and we can interpolate the virtual microphone signal.

In the virtual microphone technique, the phase and amplitude are interpolated individually. The phase and amplitude of the observed signal $x_i(\omega, \mathrm{t})$ are denoted as

$$\phi_i = \angle x_i(\omega, t) = \tan^{-1} \frac{\mathrm{Im}\big(x_i(\omega, t)\big)}{\mathrm{Re}\big(x_i(\omega, t)\big)}, \quad (1)$$

$$A_i = |x_i(\omega, t)|. \quad (2)$$

The phase ϕ_i of the virtual microphone signal is interpolated linearly as

$$\begin{aligned} \phi_{\mathrm{v}} &= \phi_1 + \alpha(\phi_2 - \phi_1) \\ &= (1 - \alpha)\phi_1 + \alpha\phi_2, \end{aligned} \quad (3)$$

where, α is the coefficient of the interpolation of a virtual microphone, which indicates that the virtual microphone is interpolated at the point obtained by internally dividing the line joining the two real microphones at the ratio α to $(1 - \alpha)$ (Fig. 1). The values of the phase are arbitrary for a natural number n in $\phi_i \pm 2\pi\mathrm{n}$. Thus, the phase of the virtual microphone is interpolated with the assumption that

$$|\phi_2 - \phi_1| \leq \pi. \quad (4)$$

The appropriate interpolation of the amplitude of the virtual microphone depends on many conditions such as the direction of arrival and reverberation. Therefore, it is quite difficult to faithfully model the real amplitude attenuation. Consequently, this method uses β-divergence for amplitude interpolation instead of some physical model. The amplitude of the virtual microphone is interpolated as

$$A_{\mathrm{v}} = \begin{cases} \exp((1 - \alpha)\log A_1 + \alpha \log A_2) & (\beta = 1) \\ \left((1 - \alpha)A_1^{\beta-1} + \alpha A_2^{\beta-1}\right)^{\frac{1}{\beta-1}} & (\text{otherwise}) \end{cases} \quad (5)$$

With the parameter β, it is possible to nonlinearly interpolate the amplitude of the virtual microphone between the amplitudes of the two real microphones. From the above, the virtual microphone signal $v(\omega, t)$ is represented as

$$v(\omega, t) = A_{\mathrm{v}} \exp(j\phi_{\mathrm{v}}). \quad (6)$$

In [12] [13], the virtual microphone signals are used as extra input signals for speech enhancement using the maximum signal-to-noise ratio (SNR) beamformer.

III. EXTRAPOLATION OF VIRTUAL MICROPHONE

In this paper, we generate extrapolated virtual microphone signals for sound image localization. The arrangement of the real and virtual microphones is shown in Fig. 2.

For the extrapolation of a virtual microphone signal, we have to check the validity of the generation method for a virtual microphone described in the previous section. For phase extrapolation, we can use the same equation as in the previous phase interpolation (see (3)).

Amplitude interpolation is difficult, and extrapolation is even more complicated. Interpolation based on β -divergence sometimes (e.g., when extrapolating a virtual microphone to a position far from the position of real microphones) (5) outputs unrealistic amplitudes such as a "negative" amplitude (which causes a complex amplitude except for β = 1) or diverges to a positive infinity. This impossible extrapolation must be avoided. Therefore, in this paper, since ITD is in the dominant infrequency range below 1.5 kHz [1][2], we use the amplitude of the real microphone closest to the virtual microphone position as the amplitude of the extrapolated virtual microphone.

$$A_{\mathrm{v}} = \begin{cases} A_1, & \alpha < 0 \\ A_2, & \alpha > 1 \end{cases} \quad (7)$$

The extrapolated virtual microphone signal is represented in the same way as in the interpolation.

$$v(\omega, t) = A_{\mathrm{v}} \exp(j\phi_{\mathrm{v}}) \quad (8)$$

In this paper, we use the estimated virtual microphone signals by extrapolation for the realignment of the real microphones for sound image localization. The estimated sound of the virtual microphone is obtained by converting the time–frequency domain signal into the time domain signal by inverse STFT.

In this research, the virtual microphone technique is applied to set a virtual microphone by extrapolating the phase difference between signals of two closely positioned real microphones. The virtual microphone is located at a distance equivalent to that between the two ears of a listener. Therefore, signals of the virtual microphone and one of the real microphones are regarded as a pair of binaural signals with adequate ITDs restored for the listener to localize sound images.

IV. EXPERIMENTS

In this experiment, using two sound sources arriving from different directions, we examined whether the virtual microphone technique that is proposed here can extrapolate the phases of two sounds individually from the detected signals of two microphones and whether the ITD is correctly restored.

4.1. Experimental conditions

In this experiment, the phases of two speech signals are extracted from the estimated virtual microphone signal and compared with those of the two sound signals observed by the real microphones.

The layout of the sound sources and real microphones is shown in Fig. 3, and other experimental conditions are shown in Table 1. M_1 and M_2 are the real microphones on the left and right, respectively. M_{v} is the virtual microphone. S_1 and S_2 are the sound sources arriving from the left side and the right side, respectively. The signals observed at M_1 and M_2 were formed by convoluting measured impulse responses into speech signals. The impulse responses in the RWCP Sound Scene Database [14] were used in this work. The impulse response signals used from the database had a reverberation time of 300 ms, a sound source distance of 2 m, and sound source directions of 10° and 170°. For the sound sources, female Japanese speech and male English speech were used.

We estimated the signal observed at M_{v} from the signals at M_1 and M_2 using the virtual microphone technique. In order to examine, it is necessary to determine the sound source in the time–frequency bin of the virtual microphone. For this purpose, we compared the power spectra of S_1 and S_2 in each time–frequency bin and constructed a binary mask that selects the larger sound source. By applying the binary mask to the estimated virtual microphone signal, the estimated signals

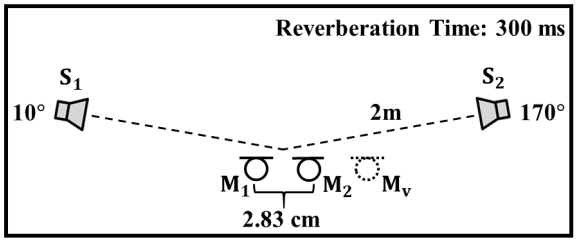

Fig. 3 Layout of sound sources and microphone in experiment

TABLE 1
EXPERIMENTAL CONDITIONS

Sampling rate	8 kHz
Real microphone interval	2.83 cm
Reverberation time	300 ms
FFT frame length	1024 samples
FFT frame shift	256 samples
Sound source 1	Female Japanese speech
Sound source 2	Male English speech

observed at M_{v} are extracted. We evaluated only time frequency bins with high power of signal. By comparing the phases of S_1 and S_2 at M_1 and the phases of S_1 and S_2 at M_{v}, we determined that the phases of S_1 and S_2 at M_{v} can be extrapolated individually, even when the signals observed at the two microphones were from two mixed sounds arriving from different directions.

To evaluate the restored ITDs, we the introduce interaural cross correlation (IACC) [2]. IACC indicates the direction of a sound image when a listener listens to the signal of M_1 with the left ear and to the signal of M_{v} with the right ear. In order to evaluate ITD on the basis of IACC, we converted S_1 and S_2 at M_{v} into time domain signals by inverse STFT. We calculated IACC from S_1 and S_2 at M_1 and M_{v} to estimate ITD as a time difference at maximum IACC.

When $\alpha = 1$, the virtual microphone corresponds to the right real microphone. When $|\alpha| = 8$, the distance between M_1 and M_{v} is equivalent to the interaural distance. When $\alpha < 0$, the virtual microphone is extrapolated to the left side of M_1 and when $\alpha > 1$ to the right side of M_2.

4.2. Results and discussion

The power spectra of signals of S_1 and S_2 for one frame are shown in Fig. 4. The experimental results at various α values are shown in Figs. 5 to 7 with (a) showing the phase difference between signals of S_1 and S_2 at M_1 and M_{v}, (b) and (c) showing the waveforms of S_1 and S_2 at M_1 and M_{v}, and (d) and (e) showing the IACC of S_1 and S_2.

Figure 4 shows the power spectrum of female Japanese speech indicated by S_1 and that of male English speech indicated by S_2. No overlap, which means "sparse" in the frequency region, is seen between the two frequency characteristics.

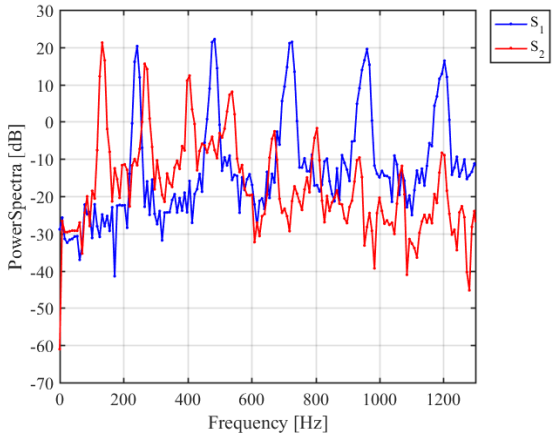

Figure. 4 Power spectra of S_1 and S_2

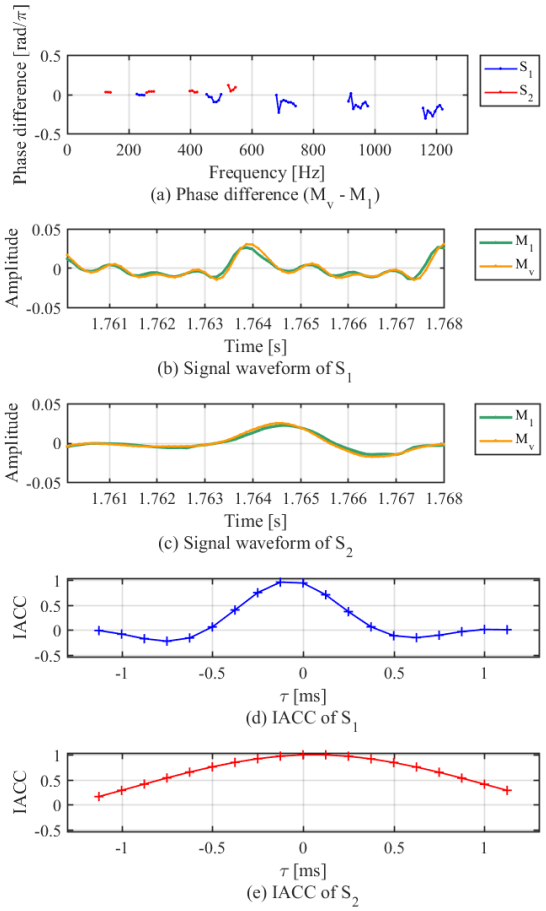

Figure. 5 Experimental result of $\alpha = 1$

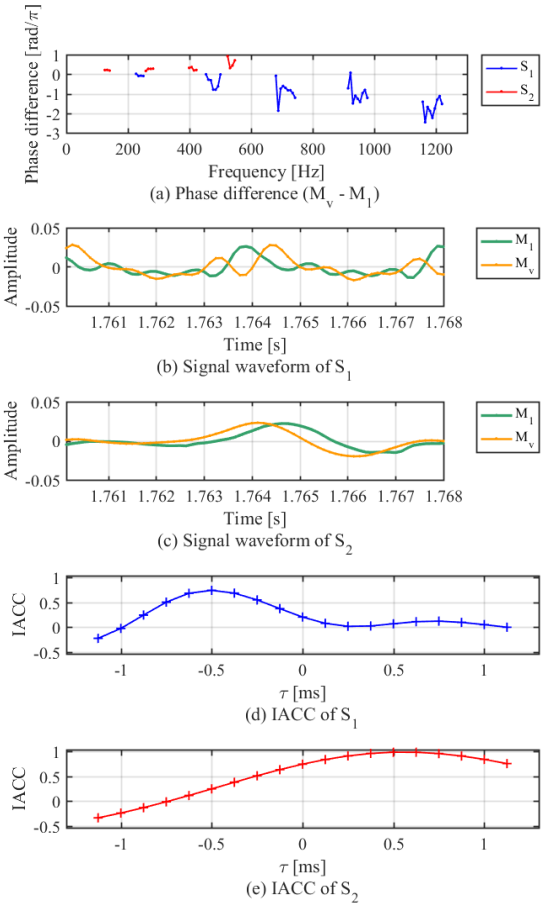

Figure. 6 Experimental result of $\alpha = 8$

Figure 5(a) shows the phase differences between signals of S_1 and S_2 at M_1 and M_v. As shown in Fig. 5(a), since S_1 is closer to M_1 than to M_v, the phase at M_1 is smaller than that at M_v (blue line). In contrast, for S_2, since it is farther from M_1 than from M_v, the phase at M_1 is larger than that at M_v (red line).

Waveforms of signals from S_1 at M_1 and M_2 are shown in Fig. 5(b). Since S_1 is closer to M_1 than to M_v, the waveform at M_1 arrival slightly earlier than that at M_v (green line). In contrast, in Fig. 5(c), since S_2 is farther from M_1 than from M_v, the arrival of the waveform at M_1 is slightly delayed (orange line).

Upon comparing phase differences between signals of S_1 and S_2 at M_1 and M_v in Fig. 5 (a) with those in Fig. 6(a), the phase differences in Fig. 6(a) are seen to be shifted by eight times those in Fig. 5(a), which correspond to the extrapolation parameter α. Similarly, the time difference between the waveforms of signals from S_1 at M_1 and that at M_v for S_1 (difference between the green and yellow lines) in Fig. 6(b) is eight times that between the waveform of signals from S_1 at M_1

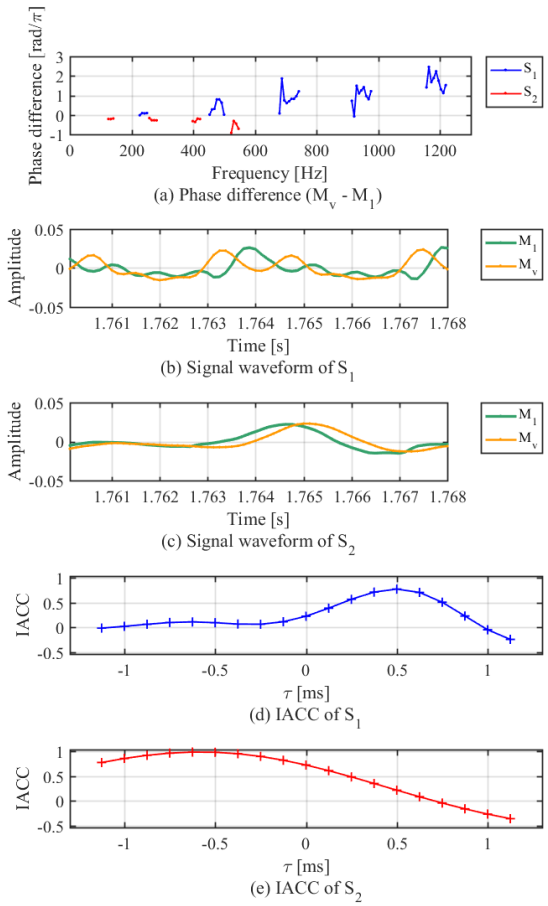

Figure. 7 Experimental result of $\alpha = -8$

and that at M_v in Fig. 5(b). This means that the distance between M_v and M_1 is eight times as long as that between M_1 and M_2 in accordance with the extrapolation coefficient α.

In Figs. 5(d) and 5(e), IACCs are close to 1 at $\tau = 0$. This means that a listener listening to the signals of M_1 and M_v in Figs. 5(b) and (c) will perceive a sound image in the median plane. In Fig. 6(d), IACC is maximum at $\tau = -500$ μs. This suggests that a listener will perceive a sound image close to the leftmost position when the listener hears the signals in Fig. 6(b). In contrast, in Fig. 6(e), IACC is maximum at $\tau = 500$ μs, a sound image for the signals in Fig. 6(c) will be localized at the rightmost position. In Figs. 7(d) and 7(e), since IACCs are opposite to those in Figs. 6(d) and 6(e), a listener will perceive sound images in directions opposite to those in Figs. 6(d) and Fig. 6(e).

V. CONCLUSION

In this paper, we applied the virtual microphone technique to realign a microphone array for restoring appropriate ITDs. A virtual microphone is assumed to be located at a distance equivalent to that between two ears of a listener by extrapolating the phase difference of signals recorded by two real microphones. A signal of the virtual microphone and that of one of the two real microphones are evaluated as a pair of binaural signals.

In the experiment, it was supposed that waveforms of signals from two sound sources in the opposite direction arrived at the two closely positioned real microphones. Furthermore, the distance between the virtual microphone and one of the real microphones is assumed eight times as long as that between the two real microphones. Time differences between signals at one of the real microphones and the virtual microphone were evaluated from the waveforms at these two microphones. Moreover, the interaural cross correlations of the signals suggest that since the ITDs are restored, sound images for the signals will be localized in directions from the viewpoint of the two real microphones.

In this method, we assume W-disjoint orthogonality of observed signals. However, in the scenes where peaks of the power spectrum of signals frequently overlap, the overlapping frequency signals have inappropriate ITDs. We will examine how to perceive the sound image when listener hear such signal by psychological evaluation experiment in the future.

This microphone realignment method requires no prior information on the number of sources and the directions of sources, and sound images for multiple sound sources from various directions are localized. Since the position of the virtual microphone is controllable by the extrapolation coefficient α, the ITD restored for the listener can be individualized by the virtual microphone technique.

ACKNOWLEDGMENT

This work was partially supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI through a Grant-in-Aid for Scientific Reserch under Grant 16H01735 and the SECOM Science and Technology Foundation.

REFERENCES

[1] B. C. J. Moore, "An Introduction to the Psychology of Hearing," 4th Edition, Academic Press, 1997.
[2] J. Blauert, "Spatial Hearing. The Psychology of Human Sound Localization," MIT Press 1996.
[3] S. Busson, R. Nicol, and B. F. G. Katz, "Subjective Investigations of the Interaural Time Difference in the Horizontal Plane," in Proc. Audio Engineering Society Convention, convention paper 6324, May 2005.
[4] A. Lindau, J. Estrella, and S. Weinzierl, "Individualization of Dynamic Binaural Synthesis by Real Time Manipulation of the ITD," in Proc. Audio Engineering Society Convention, convention paper 8088, May 2010.
[5] M. Aussal, F. Alouges, and B. F. Katz, "ITD Interpolation and Personalization for Binaural Synthesis using Spherical Harmonic," AES Conference, 04-1— 04-10 2012.
[6] F. Brinkmann, A. Lindau, S. Weinzierl, S. V. D. Par, M. Müller-Trapet, R. Opdam, and M. Vorländer, "A High Resolution and Full-Spherical Head-Related Transfer Function Database for

Different Head-Above-Torso Orientations," J. Audio Eng. Soc., Vol.65, No. 10, pp. 841 – 848, October 2017

[7] A. Härmä and C Faller, "Spatial Decomposition of Time-Frequency Regions: Subbands or Sinusoids," in Proc. Audio Engineering Society Convention, convention paper 6061, May 2004.

[8] C. Avendano and J-M Jot, "A Frequency-Domain Approach to Multichannel Upmix," J. Audio Eng. Soc., Vol. 52, No. 7/8, pp. 740-749, Jul/Aug 2002

[9] D. Barry, B. Lawlor, and E. Coyle, "Real-Time Sound Source Separation: Azimuth Discrimination and Resynthesis," in Proc. Audio Engineering Society Convention, convention paper 6258, Oct 2004.

[10] M. Combos and J. Lopez, "Interactive Enhancement of Stereo Recording Using Time-Frequency Selective Panning," AES International Conference, pp. 1 – 12, Oct 2010.

[11] N. Mae, K. Yamaoka, Y. Mitsui, M. Matsumoto, S. Makino, D. Kitamura, N. Ono, T. Yamada, and H. Saruwatari, "Ego Noise Reduction and Sound Localization Adapted to Human Ears using Hose-shaped Rescue Robot," in Proc. International Workshop on Nonlinear Circuits, Communications and Signal Processing, pp. 371-374, March 2018.

[12] H. Katahira, N. Ono, S. Miyabe, T. Yamada, and S. Makino, "Nonlinear Speech Enhancement by Virtual increase of Channels and Maximum SNR Beamformer", EURASIP Journal on Advances in Signal Processing, vol. 2016, no. 1, pp. 1-8, Jan. 2016

[13] K. Yamaoka, N. Ono, T. Yamada, and S. Makino, "Performance Evaluation of Nonlinear Speech Enhancement Based on Virtual Increase of Channels in Reverberant Environments, " in Proc. EUSIPCO, pp. 2388-2392, Aug. 2017.

[14] S. Nakamura, K. Hiyane, F. Asano, Y. Kaneda, T. Yamada, T. Nishiura, T. Kobayashi, S. Ise and H. Saruwatari, "Design and Collection of Acoustic Sound Data for Hands-Free Speech Recognition and Sound Scene Understanding," in Proc. ICME2002, Vol. 2, pp. 161-164, Aug. 2002.