Performance Evaluation of Nonlinear Speech Enhancement Based on Virtual Increase of Channels in Reverberant Environments

Kouei Yamaoka*, Shoji Makino*, Nobutaka Ono†‡, and Takeshi Yamada*
*University of Tsukuba, yamaoka@mmlab.cs.tsukuba.ac.jp, maki@tara.tsukuba.ac.jp, takeshi@cs.tsukuba.ac.jp
†National Institute of Informatics (NII)
‡SOKENDAI (The Graduate University for Advanced Studies), onono@nii.ac.jp

***Abstract*—In this paper, we evaluate the performance of a maximum signal-to-noise ratio beamformer based on a virtual increase of channels. We previously proposed a new microphone array signal processing technique, which virtually increases the number of microphones by generating extra signal channels from two real microphone signals. This technique generates a virtual observation on the assumption that the sources are W-disjoint orthogonal, which means that only one source is dominant in one time-frequency bin. However, mixed signals with a long reverberation tend to dissatisfy this assumption. In this study, we conducted experiments in a variety of reverberant environments, as well as computer simulation using image method. As a result, we confirmed that our technique contributes improving the performance in reverberant environments. We also confirmed that the longer the reverberation time, the smaller the increase in the improvement using our technique. Moreover, we present directivity patterns to confirm the behavior of a virtual increase of channels.**

I. Introduction

Microphone array signal processing is used in various techniques such as speech enhancement involving the use of beamformers and blind source separation (BSS) [1]. The speech enhancement performance using these techniques depends on the number of microphones. The performance may degrade when the number of microphones is less than the number of sound sources (underdetermined conditions).

Recently, many recording devices such as IC recorders and smartphones have become common. These devices have a small number of microphones (usually only two). For this reason, when we use these devices, speech enhancement tends to occur an underdetermined condition. Although several methods such as time-frequency masking [2], multichannel Wiener filtering [3] and the statistical modeling of observations using latent variables [4], [5] can work well in underdetermined conditions, better performance should be obtainable because they tend to contain several artificial noises such as musical noise.

As a technique for realizing high performance in underdetermined conditions, we proposed a virtual increase of channels based on *virtual microphone* signals [6]–[8]. In this technique, we create arbitrary channels of virtual microphone signals by using two channels of real microphones. Virtual microphone signals are generated as estimates of signals at a virtual microphone placed at a point where there is no real microphone. We perform microphone array signal processing using microphone signals consisting of both real and virtual microphone signals. Additionally, this technique is applicable to various types of microphone array signal processing, since we generate virtual signals in the audio signal domain, which is different from techniques in which signals are generated in the power domain [9]–[11] or a higher-order statistical domain [12], [13].

As an approach to virtual microphone signal generation, we previously proposed nonlinear interpolation using the complex logarithm spectrum of real microphone signals, which we call *complex logarithmic interpolation* [6]. Additionally, we proposed β-divergence-based nonlinear interpolation [7], [8] as a generalization of complex logarithmic interpolation. These methods assume W-disjoint orthogonality (W-DO) [2], [14]. Because of this assumption, when multiple sounds arrive, they can be regarded as a single sound. However, long reverberant environments may cause the breakdown of W-DO, and the performance of our technique in such a situation has not yet been investigated. Therefore, we study the effect of reverberation on the virtual increase of channels. In this paper, we compare the speech enhancement performance of a maximum signal-to-noise ratio (SNR) beamformer [15], [16] with the virtual increase of channels in a variety of reverberant environments.

II. Increasing channels by virtual microphone for maximum SNR beamformer

A. Increasing channels by nonlinear interpolation with β-divergence

We proposed a virtual increase of channels as a technique for creating arbitrary channels of virtual microphone signals by using two channels of real microphones [6]–[8]. By using real and generated virtual microphones, we can use a microphone array whose number of channels has been virtually increased as shown in Fig. 1.

In this technique, a microphone signal is modeled in the short-time Fourier transform (STFT) domain. Here, let $x_i(\omega, t)$ be the ith real microphone signal ($i = 1, 2$) at angular frequency ω in the tth frame. The amplitude of this signal is denoted as $A_i = |x_i(\omega, t)|$ and the phase is denoted as $\phi_i = \angle x_i(\omega, t)$. A virtual microphone signal $v(\omega, t, \alpha)$ is defined as the observation estimated at the point obtained by

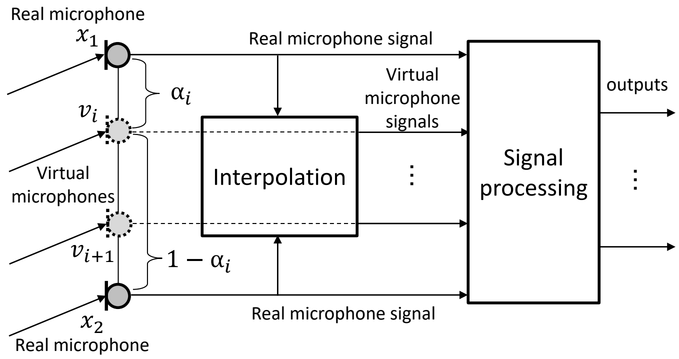


Fig. 1: Microphone array signal processing with virtual increase of channels

internally dividing the line joining two real microphones in the ratio $\alpha : (1-\alpha)$. Hereafter, when there is no need to distinguish ω, t and α, the signal is simply denoted as v.

The virtual microphone signal is obtained by a nonlinear interpolation for each time-frequency bin as follows. We derive the amplitude A_{v} that minimizes the sum $\sigma_{\boldsymbol{D}_\beta}$ of the β-divergence between the amplitude of a real microphone signal and a virtual microphone signal weighted by the virtual microphone interpolation parameter α,

$$\sigma_{\boldsymbol{D}_\beta} = (1-\alpha)\boldsymbol{D}_\beta(A_{\mathrm{v}}, A_1) + \alpha \boldsymbol{D}_\beta(A_{\mathrm{v}}, A_2), \tag{1}$$

$$A_{\mathrm{v}\beta} = \mathrm{argmin}_{A_{\mathrm{v}}} \sigma_{\boldsymbol{D}_\beta}, \tag{2}$$

where $\boldsymbol{D}_\beta(A_{\mathrm{v}}, A_i)$ is defined as

$$\boldsymbol{D}_\beta(A_{\mathrm{v}}, A_i) = \begin{cases} A_{\mathrm{v}}(\log A_{\mathrm{v}} - \log A_i) + (A_i - A_{\mathrm{v}}) & (\beta = 1), \\ \dfrac{A_{\mathrm{v}}}{A_i} - \log \dfrac{A_{\mathrm{v}}}{A_i} - 1 & (\beta = 0), \\ \dfrac{A_{\mathrm{v}}^\beta}{\beta(\beta-1)} + \dfrac{A_i^\beta}{\beta} - \dfrac{A_{\mathrm{v}} A_i^{\beta-1}}{\beta - 1} & (\text{otherwise}). \end{cases} \tag{3}$$

By differentiating $\sigma_{\boldsymbol{D}_\beta}$ with respect to A_{v} and setting it to 0, the interpolated amplitude extended using β-divergence is obtained as

$$A_{\mathrm{v}\beta} = \begin{cases} \exp\left((1-\alpha)\log A_1 + \alpha \log A_2\right) & (\beta = 1), \\ \left((1-\alpha) A_1^{\beta-1} + \alpha A_2^{\beta-1}\right)^{\frac{1}{\beta-1}} & (\text{otherwise}). \end{cases} \tag{4}$$

Note that $A_{\mathrm{v}\beta}$ is continuous at $\beta = 1$ and this interpolation is equivalent to *complex logarithmic interpolation* [6].

Phase ϕ_{v} of a virtual microphone signal v is interpolated linearly as

$$\phi_{\mathrm{v}} = (1-\alpha)\phi_1 + \alpha\phi_2, \tag{5}$$

and this interpolation requires no spatial aliasing. From the above, virtual microphone signal v is represented as

$$v = A_{\mathrm{v}\beta} \exp(j\phi_{\mathrm{v}}). \tag{6}$$

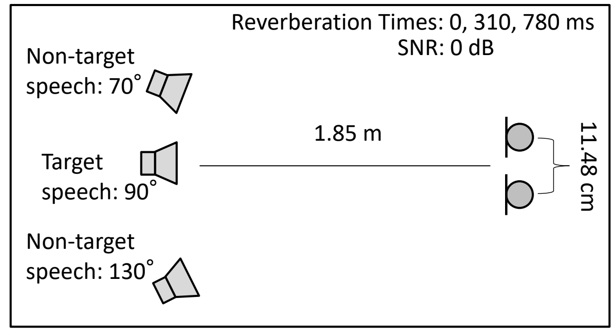


Fig. 2: Layout of sound sources and microphones in experiment

B. Maximum SNR beamformer

We apply the virtual microphone technique to a maximum SNR beamformer [15], [16], which is one of the speech enhancement techniques, to evaluate its performance. A maximum SNR beamformer requires the target-active period and target-inactive period as prior information for speech enhancement. Since this technique requires no information about the directions of sound sources, it is advantageous to application for sound sources with unknown directions.

III. Experimental evaluation of speech enhancement in reverberant environments

In this study, to evaluate the speech enhancement performance in reverberant environments, we conducted experiments using observed signals that are convolutive mixtures of measured impulse responses in a variety of reverberant environments and speech signals. Additionally, we also used the Room Impulse Response (RIR) generator [17] to simulate impulse responses. The performance was evaluated by comparing the results of two methods: speech enhancement using the maximum SNR beamformer with virtual microphone signals, and speech enhancement without virtual microphone signals (an underdetermined condition).

A. Experimental conditions

We used impulse responses in the RWCP Sound Scene Database in Real Acoustic Environments (RWCP-SSD) [18], which is a common database for evaluating speech and acoustic signal processing research in real acoustic environments. In this experiment, we used impulse responses with reverberation times of 0, 310 and 780 ms, which were measured in an anechoic chamber and variable reverberation chambers.

The layout of the sound sources and real microphones is shown in Fig. 2. The other experimental conditions are listed in Table I. We used two real microphones and generated one virtual microphone signal at their midpoint. Thus, the microphone array we used was composed of three microphones, two real microphones and one virtual microphone. Because of the long interval between the real microphones, there was spatial aliasing. We used eight samples of Japanese or English speech for the target signals, whose direction of

TABLE I: Experimental conditions

Number of real microphones	2
Number of virtual microphones	1 ($\alpha = 0.5$)
Interval between real microphones	11.48 cm
Reverberation time	0, 310, 780 ms
Input SNR	0 dB
Sampling rate	8 kHz
FFT frame length	1024 samples
FFT frame shift	256 samples
Target-active period $\lvert\Theta_T\rvert$	10 s
Target-inactive period $\lvert\Theta_I\rvert$	10 s
Speech-enhanced period	20 s

arrival (DOA) was 90°. We also used two combinations of Japanese or English speech for the non-target signals, whose DOAs were 70° and 130°. The input SNR was set to 0 dB. We used objective criteria, namely, the signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR) [19], for which higher values indicate higher performance. Here, we show the average results of SDR and SIR over the eight samples for the target signals and the two combinations for the non-target signals.

B. Results and discussion

Figure 3 shows the relationship between the speech enhancement performance and reverberation time for different values of the interpolation parameter β. Note that 'W/O virtual mic' in this figure denotes the performance of the maximum SNR beamformer without the virtual microphone and 'W/ $\beta = 0, 2, 20$' denote the performances with the virtual microphone for interpolation parameters $\beta = 0, 2$ and 20, respectively.

According to Fig. 3, the results for W/ $\beta = 0, 2, 20$ have a higher SDR and SIR than these for W/O virtual mic regardless of the reverberation time. Thus, it is confirmed that our technique contributes to improving the performance in reverberant environments. On the other hand, the longer the reverberation time, the smaller the increase in the improvement upon a virtual increase of channels. Considering these results, it can be concluded that when the W-DO requirement is not satisfied, it directly adversely affects the performance.

In a reverberant environment, the performance is improved by increasing the value of β, which controls the nonlinearity of the interpolation. In contrast, in a non-reverberant environment, the highest performance occurs when $\beta = 2$. The amplitude of the virtual microphone is interpolated linearly when $\beta = 2$ (see Eq. (4)). Essentially, we cannot obtain helpful information by linear interpolation. Nevertheless, this value of β gives the highest performance.

Regarding W-DO, Fig. 4 is a histogram showing the proportion of sources that are simultaneously active at each frequency. In this figure, we use a male-male-female combination recorded with each reverberation time. we consider that source x_i ($i = 1, 2, 3$) is active when it has a amplitude greater than $\frac{\max(|x_i|)}{10}$ for all i at each frequency [20]. If two or three sources are simultaneously active, W-DO is not satisfied. According to this figure, approximately 5% of the time-frequency bins do not satisfy W-DO in Fig. 4(a). Moreover, the percentage of time-frequency bins that do not satisfy W-DO increases with the reverberation time as shown in Figs. 4(b) and (c). For this reason, the improvement of the performance with the virtual increase of channels is decreased in the case of long reverberation.

C. Directivity patterns

Figure 5 shows directivity patterns produced by the maximum SNR beamformer in an experiment. In Fig. 5(a), the maximum SNR beamformer produced one null using two real microphones . In the frequency range of 1.5 to 4 kHz, there is spatial aliasing, so that we truncated the figure. Interestingly, according to Fig. 5(d), two nulls are created by using both one virtual microphone and two real microphones. This is the contribution due to the virtual increase of channels. However, according to Figs. 5(b) and (e), the nulls become indistinct in the case of long reverberation.

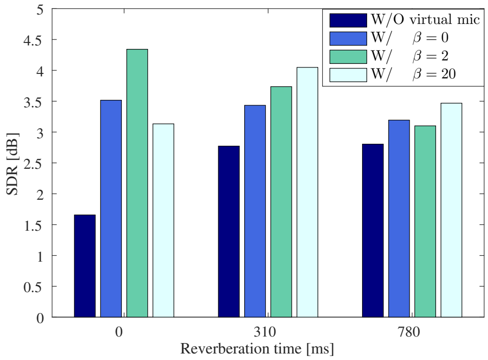
In addition to the measured impulse responses in RWCP-SSD, we also used impulse responses with reverberation times of 0, 120, 310 and 780 ms produced by the RIR generator for simulation. In this simulation, almost all the conditions were the same as those in Fig. 2 and Table I. Only the interval between the real microphones was different which we set to 4 cm to avoid spatial aliasing. Figures. 5(c) and (f) show directivity patterns obtained by the simulation. We can confirm the contribution of the virtual increase of channels more clearly than prior experiment using measured impulse responses.

IV. Conclusions

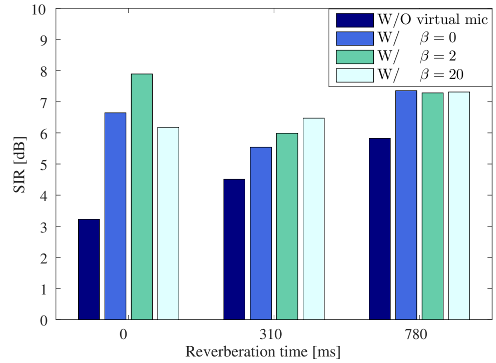
In this paper, we verified the speech enhancement performance of a maximum SNR beamformer based on a virtual increase of channels assuming W-DO, which is an important assumption for a virtual increase of channels. However, mixed signals tend not to satisfy this assumption when the reverberation time is long. Thus, we conducted experiments in a variety of reverberant environments.

As a result, we confirmed that a consistent improvement of SDR and SIR can be obtained by a virtual increase of channels even in a long-reverberation environment. However, the improvement with the virtual increase of channels decreased in the case of long reverberation. We confirmed that this is because W-DO is not satisfied by observing the number of active sources. Moreover, we showed directivity patterns to confirm the behavior of a virtual increase of channels. We expect that this significant decrease in the performance can be avoided to obtain better performance.

ISBN 978-0-9928626-7-1

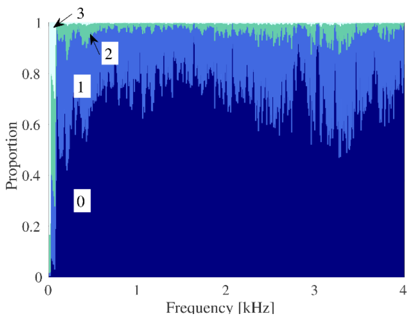


(a) SDR

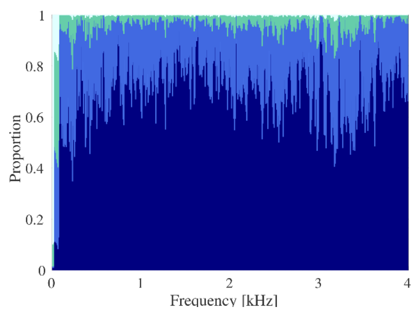


(b) SIR

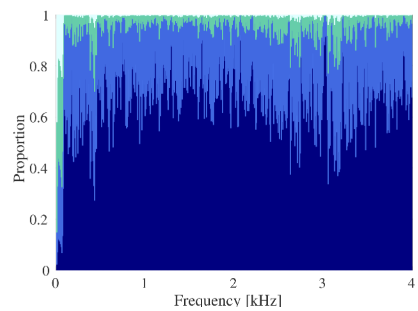
Fig. 3: Relationship between reverberation time and speech enhancement performance



(a) Reverberation time 0 ms

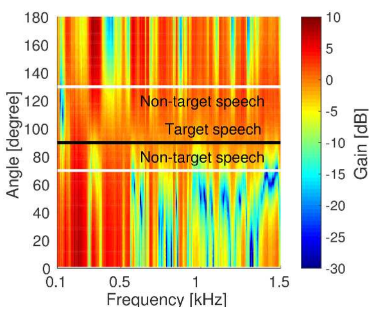


(b) Reverberation time 310 ms

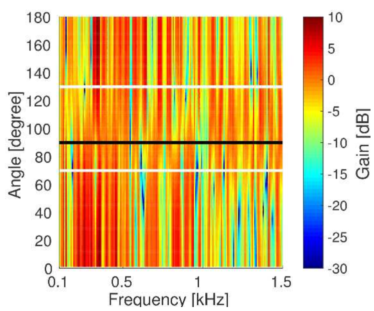


(c) Reverberation time 780 ms

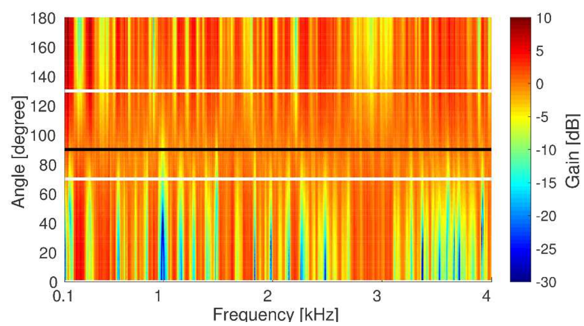
Fig. 4: Histogram showing the proportion of simultaneously active sources at each frequency



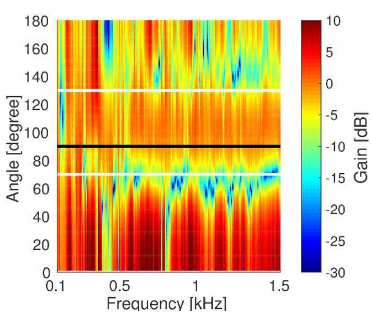
(a) W/O virtual mic ($T_{60} = 0$ ms)



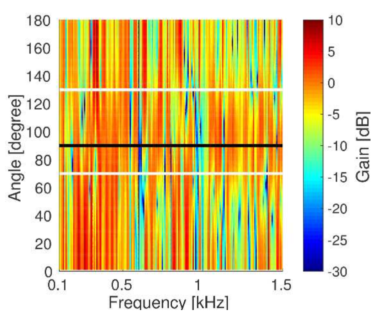
(b) W/O virtual mic ($T_{60} = 310$ ms)



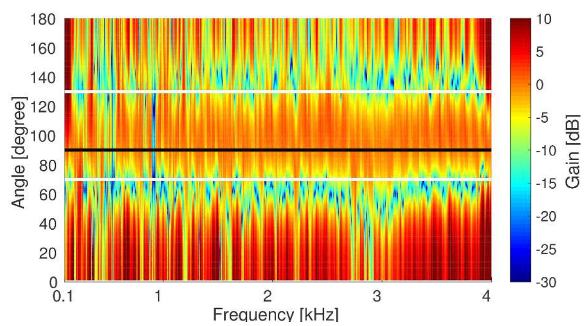
(c) W/O virtual mic ($T_{60} = 120$ ms, simulation)



(d) W/ virtual mic $\beta = 2$ ($T_{60} = 0$ ms)



(e) W/ virtual mic $\beta = 20$ ($T_{60} = 310$ ms)



(f) W/ virtual mic $\beta = 2$ ($T_{60} = 120$ ms, simulation)

Fig. 5: Directivity patterns; (a), (d) and (b), (e) used measured impulse responses with reverberation times of 0 and 310 ms, respectively, (c), (f) used calculated impulse responses with a reverberation time of 120 ms

ISBN 978-0-9928626-7-1

REFERENCES

[1] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*, Springer, 2007.

[2] O. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, pp. 1830–1847, 2004.

[3] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[4] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," *Proc. WASPAA*, pp. 147–150, 2007.

[5] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[6] H. Katahira, N. Ono, S. Miyabe, T. Yamada, and S. Makino, "Virtually increasing microphone array elements by interpolation in complex-logarithmic domain," *Proc. EUSIPCO*, pp. 1–5, Sept. 2013.

[7] H. Katahira, N. Ono, S. Miyabe, T. Yamada, and S. Makino, "Generalized amplitude interpolation by β-divergence for virtual microphone array," *Proc. IWAENC*, pp. 150–154, Sept. 2014.

[8] 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.

[9] H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Speech enhancement using nonlinear microphone array based on complementary beamforming," *IEICE Trans. on Fundamentals*, vol. E82-A(8), pp. 1501–1510, 1999.

[10] S. Miyabe, B. H. (Fred) Juang, H. Saruwatari, and K. Shikano, "Analytical solution of nonlinear microphone array based on complementary beamforming," *Proc. IWAENC*, pp. 1–4, 2008.

[11] Y. Hioka and T. Betlehem, "Under-determined source separation based on power spectral density estimated using cylindrical mode beamforming," *Proc. WASPAA*, pp. 1–4, 2013.

[12] P. Chevalier, A. Ferréol, and L. Albera, "High-resolution direction finding from higher order statistics: The 2q-MUSIC algorithm," *IEEE Trans. on Signal Processing*, vol. 53, no. 4, pp. 2986–2997, 2006.

[13] Y. Sugimoto, S. Miyabe, T. Yamada, S. Makino, and B. H. (Fred) Juang, "Employing moments of multiple high orders for high-resolution underdetermined DOA estimation based on music," *Proc. WASPAA*, pp. 1–4, 2013.

[14] A. Jourjine, S. Rickard, and O. Yılmaz, "Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures," *Proc. ICASSP*, pp. 2985–2988, 2000.

[15] H. L. Van Trees, *Optinum Array Processing*, John Wiley & Sons, 2002.

[16] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," *Proc. ICASSP*, vol. I, pp. 41–45, 2007.

[17] E. A. P. Habets, "Room impulse response (RIR) generator," Available at: https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator, Oct. 2008.

[18] 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," *Proc. ICME '02*, vol. 2, pp. 161–164, Aug. 2002.

[19] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[20] A. Blin, S. Araki, and S. Makino, "Underdetermined blind separation of convolutive mixtures of speech using time-frequency mask and mixing matrix estimation," *IEICE Trans. on Fundamentals*, vol. 88-A, no. 7, pp. 1693–1700, 2005.

ISBN 978-0-9928626-7-1 © EURASIP 2017