BLIND COMPENSATION OF INTER-CHANNEL SAMPLING FREQUENCY MISMATCH WITH MAXIMUM LIKELIHOOD ESTIMATION IN STFT DOMAIN

†*Shigeki Miyabe,* ‡*Nobutaka Ono,* †*Shoji Makino*

†Life Science Center of Tsukuba Advanced Research Alliance, University of Tsukuba
{miyabe,maki}@tara.tsukuba.ac.jp
‡Principles of Informatics Research Division, National Institute of Informatics
onono@nii.ac.jp

ABSTRACT

This paper proposes a novel blind compensation of sampling frequency mismatch for asynchronous microphone array. Digital signals simultaneously observed by different recording devices have drift of the time differences between the observation channels because of the sampling frequency mismatch among the devices. Based on the model that such the time difference is constant within each time frame, but varies proportional to the time frame index, the effect of the sampling frequency mismatch can be compensated in the short-time Fourier transform domain by the linear phase shift. By assuming the sources are motionless and stationary, a likelihood of the sampling frequency mismatch is formulated. The maximum likelihood estimation is obtained effectively by a golden section search.

Index Terms— Asynchronous microphone array, sampling frequency, maximum likelihood estimation, blind source separation

1. INTRODUCTION

Asynchronous microphone array is a concept of microphone array signal processing where audio signals recorded by separated multiple devices are used as the multichannel signal for array signal processing [1]. One typical application is speech enhancement of audio signals recorded by multiple mobile recording devices. The advantage of asynchronous microphone array is its freedom in the choice of the recording devices for many-channel recording, and it requires no large-scale recording devices such as special microphones for microphone arrays or many-channel analog-to-digital converters (ADCs). However, the asynchronous channels brings many additional issues which are not treated conventionally in microphone array signal processing. For example, the array geometry is naturally unknown [2, 3, 4, 5], the recording devices have different unknown gains [6], each device starts recording independently [4, 5], and the sampling frequencies are not common among the observation channels [1, 7, 8, 9].

Among many issues of asynchronous microphone array described above, one of the most important issues is the mismatch of sampling frequencies. Sampling frequencies of ADCs have bias of the order of 10 ppm (parts per million, 10^{-6}) mainly caused by individual variability of the quartz in their clock generators. Thus mismatch of sampling frequencies is inevitable without synchronization. The difference in the unit lengths of samples causes drift of time difference among observed digital signals in different channels. Since most of array signal processing methods assume that locations of sound sources have unique time differences of arrival (TDOAs) among observation channels, even a sample of change in the TDOAs is considerably large for array signal processing. Thus bias of the order of 10 ppm is significant for array signal processing.

In this paper we propose a blind compensation of sampling frequency mismatches between channels. First, modeling the drift within a frame to be constant time shift proportional to the time of the frame as modeled similarly in [9], we compensate the sampling frequency mismatch in the short-time Fourier transform (STFT) domain by the linear phase shift. Second, assuming the sources are motionless, we introduce a likelihood of sampling frequency mismatch estimate. Since the drift of time differences appears as if the sources are moving, we derive the likelihood function to evaluate the spatial stationarity of the compensated observation. Although the maximum likelihood estimation cannot be solved analytically, the optimization is solved efficiently by combination of coarse discretized search and fine golden section search employing the experimentally found property that the likelihood function is usually unimodal locally around the global maximum. The experimental results show that the proposed method effectively recovers performance of blind source separation (BSS) [10]. Note that we focus on the compensation of the sampling frequency mismatch between two channels, but it can be extended to arbitrary number of channels by fixing one reference channel and fitting the others to the reference.

2. MODEL OF SAMPLING FREQUENCY MISMATCH

2.1. Time domain model of sampling frequency mismatch

Suppose sound pressures $\mathrm{x}_1(t)$ and $\mathrm{x}_2(t)$ on two microphones are sampled by different ADCs as $x_1(n)$ and $x_2(n)$, where t denotes the continuous time and n gives the discrete time. Also suppose the sampling frequency of $x_1(n)$ is f_s, and that of $x_2(n)$ is $(1+\epsilon) f_\mathrm{s}$ with a dimensionless number ϵ. This paper assumes that the ADCs have the common nominal sampling frequencies and $|\epsilon| \ll 1$. Then the relations between $x_i(n)$ and $\mathrm{x}_i(t)$ for $i = 1, 2$ are given by

$$x_1(n) = \mathrm{x}_1\left(\frac{n}{f_\mathrm{s}}\right), \tag{1}$$

$$x_2(n) = \mathrm{x}_2\left(\frac{n}{(1+\epsilon) f_\mathrm{s}} + T_{21}\right), \tag{2}$$

where the origin of the continuous time t is defined as the starting time of the sampling of $x_1(n)$, and T_{21} is the time when the sampling of $x_2(n)$ starts. The discrete times n_1, n_2 of these two

This work is supported by an NII Grand Challenge project grant.

sampled signals $x_1(n_1)$, $x_2(n_2)$ correspond to the same continuous time t are given by

$$n_2 = (1+\epsilon)(n_1 - f_s T_{21}). \tag{3}$$

Note that hereafter we use the notation n_1 and n_2 to denote the pair of the discrete time corresponding to the same time, and simply use the notation n when we don't need to consider such the correspondence. Considering the integer-valued discrete time sample n_1, the corresponding discrete time n_2 is generally non-integer. Thus, to compensate the sampling of $x_2(n)$ and obtain its modified version $\hat{x}_2(n)$ synchronized with $x_1(n)$ as

$$\hat{x}_2(n) = \mathrm{x}_2\left(\frac{n}{f_s}\right), \tag{4}$$

we need an infinite convolution of a sinc function assuming the infinite samples of $x_2(n)$ are available;

$$\begin{aligned}\hat{x}_2(n) &= x_2\left((1+\epsilon)(n - f_s T_{21})\right) \\ &= \sum_{n'=-\infty}^{\infty} \mathrm{sinc}\left((1+\epsilon)(n - f_s T_{21}) - n'\right) x_2\left(n'\right). \end{aligned} \tag{5}$$

This convolution is complicated even if the convolution is truncated at a certain finite length, and we need some simpler approximation to formulate effective estimation of the sampling frequency mismatch.

2.2. Modeling of sampling frequency mismatch in STFT domain

Since array signal processing is usually conducted in the STFT domain, we discuss an approximation of the sampling frequency mismatch compensation in the STFT domain. Before proceeding to the discussion of STFT, we analyze how the sampling mismatch appear inside short-time frames. We consider the frame analysis of $x_1(n_1)$ with the frame length M and the central sample of the frame is denoted by m, and the range of n_1 inside the frame is limited to $m - L/2 \le n_1 \le m + L/2 - 1$. From Eq. (3), the discrete times n_1 and n_2 of the two observation channels satisfy the condition;

$$\begin{aligned} n_2 &= (1+\epsilon)(n_1 - m) + (1+\epsilon)(m - f_s T_{21}) \\ \Leftrightarrow (n_2 - m) &= (1+\epsilon)(n_1 - m) + \epsilon m - (1+\epsilon) f_s T_{21}. \end{aligned} \tag{6}$$

Since $|\epsilon(n_1 - m)|$ is typically close to zero, we regard it to be zero and obtain the following approximation model.

$$\begin{aligned}(n_2 - m) &\approx (n_1 - m) + \epsilon m - (1+\epsilon) f_s T_{21} \\ &= (n_1 - m) + \tau_{21}(m;\epsilon), \end{aligned} \tag{7}$$

where the drift inside a frame is ignored and the effect of the mismatch is regarded as a steady time shift $\tau_{21}(m;\epsilon)$ given by

$$\tau_{21}(m;\epsilon) = \epsilon(m - M), \tag{8}$$

$$M = \frac{1+\epsilon}{\epsilon} f_s T_{21}, \tag{9}$$

where M denotes the discrete time when n_1 and n_2 coincide, i.e., when $n_1 = M$, $n_2 = M$. Hereafter we call M as *offset origin*.

If the length of the shift is much smaller than the frame size as

$$|\tau_{21}(m;\epsilon)| = |\epsilon(m - M)| \ll L, \tag{10}$$

for m of all the frames to be analyzed, the compensation of the time shift $\tau_{21}(m;\epsilon)$ can be approximated well as the time shift inside the frame. The simplest approximation to compensate for the non-integral time shift $\tau_{21}(m;\epsilon)$ is to filter the linear phase in the STFT domain. First we apply analogous frame analyses to $x_i(n)$ and obtain the STFT-domain signal $X_i(k,m)$ for $i = 1, 2$ as

$$X_i(k,m) = \sum_{l=0}^{L-1} w(l)\, x_i\left(l + m - \frac{L}{2}\right) \exp\left(-\frac{2\pi \jmath k l}{L}\right), \tag{11}$$

where $w(l)$ is the window function and $k = -L/2+1, \ldots, L/2$ denotes the discrete frequency index. Compensation for the effect of the time shift $\tau(m;\epsilon)$ in $X_2(k,m)$ to obtain the approximation $\hat{X}_2(k,m;\epsilon)$ with the synchronization to $X_1(k,m)$ is given by

$$\hat{X}_2(k,m;\epsilon) = X_2(k,m) \exp\left(\frac{2\pi \jmath k \tau_{21}(k,m)}{L}\right). \tag{12}$$

This STFT-domain approximation is advantageous in its simple operation. In the following section we discuss an iterative optimization of the sampling frequency mismatch ϵ in the STFT domain based on this model. The compensation of Eq. (12) does not require the repetition of the frame and STFT analysis in each iteration. However, the linear phase shift in the STFT domain is just an approximation of the time shift in the time domain, and the approximation error is large when the condition in Eq. (10) is not satisfied. To satisfy this condition, both $|m - M|$ and $|\epsilon m|$ have to be small. The former can be satisfied to some extent by the correlation analysis in the following section, but the latter cannot be solved when the absolute value $|\epsilon|$ of the sampling frequency mismatch is large and the length of the observed signal is long. We do not focus on the treatment of such cases in this paper, but in such cases, the frame analysis of $x_2(n)$ should not be analogous to that of $x_1(n)$ as in Eq. (11), but the central samples of the frames should be modified taking the effect of the time shift $\tau(m;\epsilon)$ into account.

3. MAXIMUM LIKELIHOOD ESTIMATION OF SAMPLING FREQUENCY MISMATCH ASSUMING SPATIAL STATIONARITY

3.1. Rough compensation of whole-sample offset

Blind identification of the recording time offset T_{21} in Eq. (2) is not easy only with the observed signals $x_1(n)$ and $x_2(n)$. However, accurate compensation of T_{21} is unnecessary and small constant time rags among channels is accepted in some classes of array signal processing, such as BSS [10] and maximum signal-to-noise beamformer [11]. Thus we compensate for the recording time offset T_{21} roughly.

Usually the sound pressures $\mathrm{x}_1(t)$ and $\mathrm{x}_2(t)$ at the microphones are highly correlated. Also we are assuming that the sampling frequency mismatch is small and $|\epsilon| \ll 1$. Thus the observed signals $x_1(n)$ for $n = 0, \ldots, N_1 - 1$ and $x_2(n)$ for $n = 0, \ldots, N_2 - 1$ have high correlation even without the compensation of the sampling frequency mismatch. Thus we find the whole-sample offset δ_{21} given as delay in $x_2(n_2)$ to maximize the correlation as

$$\delta_{21} = \underset{-N_2 < \delta < N_1}{\arg\max} \sum_{n=\max(0,\delta)}^{\min(N_1, N_2+\delta)-1} x_1(n)\, x_2(n-\delta), \tag{13}$$

and to compensate the whole-sample offset δ_{21}, we give delay to $x_2(n)$ as

$$x_2(n) \leftarrow x_2(n - \delta_{21}). \tag{14}$$

Also, the offset origin in Eq. (9) must be in the middle of the overlap of $x_1(n)$ and $x_2(n-\delta_{21})$. Thus we give the following estimate of M:

$$M \leftarrow \left\lfloor \frac{\min(N_1 - \delta_{21}, N_2) - \max(0, \delta_{21}) - 1}{2} \right\rfloor, \quad (15)$$

where $\lfloor \cdot \rfloor$ denotes rounding down. With these whole-sample offset compensation and the offset origin estimation, the term $|m - M|$ in Eq. (10) is made as small as possible.

3.2. Likelihood to evaluate estimation of sampling frequency mismatch assuming spatial stationarity

The drift caused by the sampling frequency mismatch makes the TDOAs of each sound source change slowly according to the proceed of the time. Thus if the movements of sources are not large, the compensation of the sampling frequency mismatch can be evaluated with how static the TDOAs are. Also by assuming that the sources are stationary, spatial stationarity can be the measure of the sampling frequency mismatch. In this section we derive a likelihood of the sampling frequency mismatch according to these assumptions.

We assume all the sources do not move and their amplitudes are considered to be stationary on a long-term basis. We also assume that in the STFT domain the amplitudes follow the zero-mean normal distributions. Under these assumptions, the compensated observed signal $\hat{\mathbf{X}}(k, m; \epsilon)$ in the vector notation, given by

$$\hat{\mathbf{X}}(k, m; \epsilon) = \left[X_1(k, m), \hat{X}_2(k, m; \epsilon) \right]^{\mathsf{T}}, \quad (16)$$

is regarded to be stationary and follow the zero-mean multivariate normal distribution if the sampling frequency mismatch ϵ is estimated accurately and made stationary. Thus accurate estimation of ϵ tends to maximize the following log likelihood function $J(\mathbf{V}, \epsilon)$ to evaluate the fit with the zero-mean multivariate normal distribution:

$$J(\mathbf{V}, \epsilon) = \sum_{k,m} \big(-\log \pi^2 - \log\det \mathbf{V}(k) - \hat{\mathbf{X}}(k, m; \epsilon)^{\mathsf{H}} \mathbf{V}(k)^{-1} \hat{\mathbf{X}}(k, m; \epsilon) \big), \quad (17)$$

where $\mathbf{V}$ denotes the group of all the covariance matrices $\mathbf{V}(k)$, i.e., $\{\mathbf{V}(k) \,|k = -L/2+1, \ldots, L/2\}$, which is another parameter to be optimized in the maximum likelihood estimation. The covariance matrix $\mathbf{V}(k)$ is given by the following sample estimate:

$$\mathbf{V}(k) \leftarrow \frac{1}{|\forall m|} \sum_{\forall m} \hat{\mathbf{X}}(k, m; \epsilon) \hat{\mathbf{X}}(k, m; \epsilon)^{\mathsf{H}}, \quad (18)$$

where $|\forall m|$ denotes the number of the frames. By omitting the constants, the simplified version $J(\epsilon)$ of the log likelihood function $J(\mathbf{V}, \epsilon)$ is given by

$$J(\epsilon) = -\sum_{k} \log\det \sum_{\forall m} \hat{\mathbf{X}}(k, m; \epsilon) \hat{\mathbf{X}}(k, m; \epsilon)^{\mathsf{H}}. \quad (19)$$

Since the estimate of ϵ to maximize the likelihood cannot be obtained analytically, its efficient search is described in the following.

3.3. Efficient optimization of maximum likelihood estimate by golden section search

Since the parameter to be optimized in the maximum likelihood estimation is only the sampling frequency mismatch ϵ, we can use the golden section search, a representative line search method. It finds a minimum or maximum of a unimodal function by reducing the search range iteratively. As an example shown in Fig. 1, usually the log likelihood function $J(\epsilon)$ given by Eqs. (16), (19) is usually unimodal locally around the global maximum. Thus after specifying the unimodal range including the global maximum, the golden section search can be utilized.

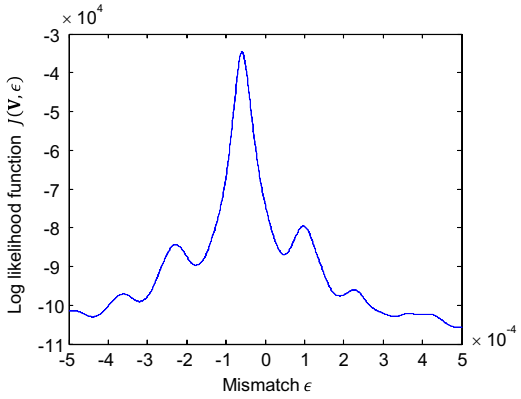

Fig. 1. An examples of the log likelihood function $J(\epsilon)$. Local concavity around the global maximum can be seen.

To find the unimodal range around the maximum, we discretize ϵ roughly and select the discretized value to maximize the log likelihood function $J(\epsilon)$. Suppose we discretize ϵ into D samples ϵ_d uniformly in the range of $[-E, E]$ as

$$\epsilon_d = -E + \frac{2dE}{D-1}, \quad d = 0, 1, \ldots, D-1. \quad (20)$$

Then we compare all the values of $J(\epsilon_i)$ to find the maximum as

$$d^* = \underset{d=0,\ldots,D-1}{\arg\max}\ J(\epsilon_d). \quad (21)$$

The range parameter E can be decided easily by considering the possible range of ϵ. Since the sampling frequency mismatch normally takes the value in the order of 10^{-5}, E can be set around 10^{-4} or larger. Appropriate setting of D depends on the range parameter E, and $2E/(D-1) < 10^{-4}$ must be satisfied according to the curve in Fig. 1.

After the search of the discretized values and obtain the rough estimate ϵ_{d^*}, the golden section search is applied to reduce the search range. We show the algorithm in Table 1. The initial search range is determined by $[\epsilon_{d^*-1}, \epsilon_{d^*+1}]$, and the iteration continues until the range shrinks to the desired resolution $\rho\ (> 0)$.

4. EXPERIMENTAL EVALUATION

To confirm the effectiveness of the proposed blind sampling frequency mismatch compensation, we gave artificial sampling frequency mismatch to observation of two speakers' speech with two microphones, and evaluated the accuracy of the sampling frequency mismatch compensation and its contribution to BSS.

4.1. Experimental conditions

The observed signals are made by convolution of measured impulse responses and speech signals, which are made by concatenation of word utterances chosen from ATR Japanese speech database [12].

Table 1. The algorithm to search the maximum of $J(\epsilon)$

Definition and initialization:
$\varphi = \frac{\sqrt{5}-1}{2}$
Set $a \leftarrow \epsilon_{d^*-1}$, $b \leftarrow \epsilon_{d^*+1}$
Step 1:
Set $p \leftarrow b - \varphi(b-a)$, $q \leftarrow a + \varphi(b-a)$
Calculate $J(p)$ and $J(q)$ according to Eqs. (16), (19)
Step 2:
If $J(p) \leq J(q)$
Set $a \leftarrow p$, $p \leftarrow q$, $q \leftarrow a + \varphi(b-a)$
Else
Set $b \leftarrow q$, $q \leftarrow p$, $p \leftarrow b - \varphi(b-a)$
End if
Step 3:
If $b - a > \rho$
Go back to Step 1
Else
Obtain the result $\epsilon \leftarrow \frac{a+b}{2}$
Terminate the algorithm
End if

Table 2. Experimental conditions

Signal length	3, 5, 10, 20, 30 seconds
Reverberation time	T_{60} of 130 ms
Frame length L	4,096 samples
Frame shift M	2,048 samples
Source distance	1.5 m
Source directions	[-50°, 30°], [-60°, -10°]
Microphone spacing	2 cm
Discretization search range E	5×10^{-4}
discretization division D	10
Golden section search resolution ρ	10^{-9}

We evaluated all the 12 combinations of two speakers from two male and two female speakers. The original sampling frequency of the observation is 16,000 kHz, and to one channel we gave modifications of sampling frequency of ± 0.5, ± 1, ± 1.5 Hz. Those correspond to the sampling frequency mismatch of ± 31.25, ± 62.5, ± 93.75 ppm, which are realistic as practical bias of sampling frequencies. To generate the artificial sampling frequency mismatch, we used resampling with the polyphase filters of 100 taps. We used auxiliary-function-based independent vector analysis [13] to conduct BSS. Other conditions are listed in Table 2.

4.2. Accuracy of sampling frequency mismatch estimation

The root mean squared errors (RMSEs) of the estimation of sampling frequency mismatches for different signal lengths are listed in Table 3. We can see that our estimation algorithm works appropriately even with short observed signals, and the accuracy improves according to the increase of the length of the observed signals.

4.3. Contribution to BSS

We compared the separation performances of different conditions of sampling frequency mismatch compensation to evaluate the contribution of the proposed method to recover the performance of

Table 3. Estimation errors for sampling frequency mismatch

Signal length [s]	3	5	10	20	30
RMSE [ppm]	2.2	1.4	0.43	0.19	0.086

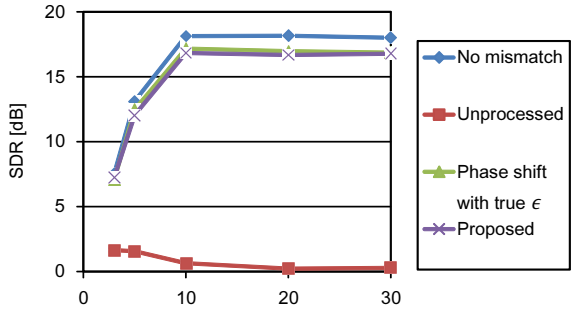

Fig. 2. Signal-to-distortion ratios (SDRs) of BSS performances for different signal conditions. No mismatch is the observation without sampling frequency mismatch. Unprocessed shows the observation with the sampling frequency mismatch not compensated. Phase shift with true ϵ shows the sampling frequency mismatch compensation by the linear phase shift with the true sampling frequency mismatch ϵ given manually. Proposed shows the proposed blind sampling frequency mismatch compensation by the linear phase shift with the maximum likelihood estimate.

BSS. We evaluated the source separation performance with signal-to-distortion ratio (SDR) [14] of the first channel as shown in Fig. 2. Since the SDRs of the unprocessed signal is very low, we can see that BSS is hard with the sampling frequency mismatches of these conditions, and the compensation of the sampling frequency mismatches is necessary. The result of the linear phase shift with manually-given true sampling frequency mismatch is only 2 dB lower than that without the sampling frequency mismatch. Thus we can say that the effect of the error of the linear phase shift on BSS is small. The blind compensation by the proposed method has similarly high SDR to the manual compensation, and the estimation error is not significant for BSS. Therefore it is confirmed that the proposed method can effectively recover the degradation of BSS caused by sampling frequency mismatch.

5. CONCLUSIONS

In this paper we proposed a novel blind compensation of sampling frequency mismatch between observation channels of asynchronous microphone array. We introduced an approximation model that the drift caused by the sampling frequency mismatch is regarded as a static time shift proportional to the sampling frequency mismatch. According to this model, we proposed the sampling frequency mismatch compensation in the STFT domain as the linear phase shift. Also we proposed the blind estimation of optimal linear phase shift. Assuming the sources are spatially stationary, we introduced a probabilistic model of the compensated STFT signals to follow the stationary zero-mean multivariate normal distribution. The maximum likelihood estimation of the phase shift is efficiently solved by golden section search. By evaluating BSS of the compensated observation, we confirmed that the maximum likelihood linear phase shift effectively recovers the degradation of source separation performance caused by sampling frequency mismatch.

6. REFERENCES

[1] Z. Liu, "Sound source separation with distributed microphone arrays in the presence of clock synchronization errors," *Proc. IWAENC*, 2008.

[2] V. C. Raykar, I. V. Kozintsev and R. Lienhart, "Position calibration of microphones and loudspeakers in distributed computing platforms," *IEEE Trans. Speech Audio Process.,* vol. 13, no. 1, pp. 70–83, 2005.

[3] M. Chen, Z. Liu, L.-W. He and P. Chou, "Energy-based position estimation of microphones and speakers for ad hoc microphone arrays," *Proc. WASPAA,* pp. 22–25, 2007.

[4] N. Ono, H. Kohno, N. Ito and S. Sagayama, "Blind alignment of asynchronously recorded signals for distributed microphone array," *Proc. WASPAA*, pp. 161–164, 2009.

[5] K. Hasegawa, N. Ono, S. Miyabe and S. Sagayama, "Blind estimation of locations and time offsets for distributed recording devices," *Proc. LVA/ICA*, pp. 57–64, 2010.

[6] Z. Liu, Z. Zhang, L.-W. He and P. Chou, "Energy-based sound source localization and gain normalization for ad hoc microphone arrays," *Proc. ICASSP,* pp. 761-764, 2007.

[7] R. Lienhart, I. Kozintsev, S. Wehr and M. Yeung, "On the importance of exact synchronization for distributed audio processing," *Proc. ICASSP*, pp. 840–843, 2003.

[8] E. Robledo-Arnuncio, T. S. Wada and B.-H. Juang, "On dealing with sampling rate mismatches in blind source separation and acoustic echo cancellation," *Proc. WASPAA*, pp. 21–24, 2007.

[9] S. Markovich-Golan, S. Gannot, I. Cohen, "Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming," *Proc. IWAENC*, 2012.

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

[11] S. Araki, H. Sawada and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," *Proc. ICASSP,* vol. 1, pp. 41–44, 2007.

[12] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.

[13] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," *Proc. WASPAA*, pp. 189–192, 2011.

[14] E. Vincent, H. Sawada, P. Bofill, S. Makino and J. P. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," Proc. ICA, pp. 552–559, 2007.