# Speech enhancement with ad-hoc microphone array using single source activity

Ryutaro Sakanashi*, Nobutaka Ono†, Shigeki Miyabe‡, Takeshi Yamada§ and Shoji Makino¶
*‡§¶Graduate School of Systems and Information Engineering, University of Tsukuba, Japan
†National Institute of Informatics / School of Multidisciplinary Sciences,
The Graduate University for Advanced Studies (SOKENDAI) , Japan
E-mail: *sakanashi@mmlab.cs.tsukuba.ac.jp, †onono@nii.ac.jp,
{‡miyabe, ¶maki}@tara.tsukuba.ac.jp, §takeshi@cs.tsukuba.ac.jp

***Abstract*—In this paper, we propose a method for synchronizing asynchronous channels in an ad-hoc microphone array based on single source activity for speech enhancement. An ad-hoc microphone array can include multiple recording devices, which do not communicate with each other. Therefore, their synchronization is a significant issue when using the conventional microphone array technique. We here assume that we know two or more segments (typically the beginning and the end of the recording) where only the sound source is active. Based on this situation, we compensate for the difference between the start and end of the recording and the sampling frequency mismatch. We also describe experimental results for speech enhancement with a maximum SNR beamformer.**

## I. Introduction

A microphone array using multiple microphones can perform various types of signal processing by obtaining spatial information from the phase difference of the sound waves that reach the microphones. There has been rapid progress in research on the application of this technology to hands-free speech recognition and the understanding of the sound environment by computers in real environments. For example, a beamformer enhances target speech by controlling directional characteristics. Also, blind source separation extracts speech sources from the mixed observation of multiple sources without prior information. Generally, these microphone array signal processing techniques assume that the microphone elements are placed regularly and the recording channels are synchronized properly with unified multichannel A/D converters, and such requirements result in the limited applicability of microphone arrays because of the need to use special expensive equipment.

To extend the application of microphone array signal processing, increasing attention has been paid to ad-hoc microphone arrays, which use multiple independent recording devices for multichannel speech signal processing. The advantage of the ad-hoc microphone array is that it gives us freedom when choosing recording devices for many-channel recording, and it requires no large-scale recording devices such as special microphones or many-channel analog-to-digital converters (ADCs). However, asynchronous channels have many additional issues that are not dealt with in conventional microphone array signal processing. For example, the array geometry is unknown, the recording devices have different unknown gains, and each device starts recording independently. In particular, the sampling frequencies are not common to all the observation channels because of independent A/D converters, and sampling frequency mismatches are inevitable. The difference between the unit lengths of samples causes the time difference between observed digital signals in different channels to drift. Since most of array signal processing methods assume that the locations of sound sources have unique time differences of arrival (TDOAs) among observation channels, even a sample of change in the TDOAs is very large for array signal processing.

Several studies have tried to deal with this problem. On the assumption that there was no sampling frequency mismatch, some authors proposed blind alignment to estimate the recording start time and the positions of microphones and sources simultaneously. Robledo *et al.* examined compensation of the sampling mismatch by resampling with interpolation. For blind estimation of sampling mismatch, Liu *et al.* utilized the correlation of an amplitude spectrogram. Markovich *et al.* proposed the semi-blind compensation of sampling mismatch with given speech absence information. Recently, we proposed the accurate blind compensation of sampling mismatch assuming stationarity of observation.

In this paper, we propose a user-guided speech enhancement framework in an ad-hoc microphone array scenario assuming that two short intervals of target speech activity are specified by the user. By estimating the inter-channel time difference of the specified intervals, the identification of the sampling mismatch is no longer a blind estimation problem. The intervals are also used for the adaptation of a maximum signal-to-noise ratio (SNR) beamformer for speech enhancement. Although the well-used approach to adaptation with a steering vector is not easy in the distributed microphone array scenario where the positions of speakers and microphones are unknown, a maximum SNR beamformer optimizes its directivity using speech activity information without the steering vector. Experimental results show that our proposed method enhances the target speech successfully by employing the multichannel attribute of an ad-hoc microphone array.

## II. Time domain model of asynchronous recording

First, we discuss the formulation of the drift in asynchronous recording. Although we limit the discussion to the sampling frequency mismatch between two channels in this paper, it is easy to extend it to three or more channels by fitting sampling frequency of all channels to one specific channel. Suppose that sound pressures $\mathrm{x}_1(t)$ and $\mathrm{x}_2(t)$ on two microphones are sampled by different ADCs as $x_1(n_1)$ and $x_2(n_2)$, where $t$ denotes continuous time and $n$ gives the discrete time. Also suppose that the sampling frequency of $x_1(n_1)$ is $f_\mathrm{s}$, and that of $x_2(n_2)$ is $(1+\epsilon)f_\mathrm{s}$ with a dimensionless number $\epsilon$. This paper assumes that the ADCs have 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

$$x_1(n_1) = \mathrm{x}_1\left(\frac{n_1}{f_\mathrm{s}}\right) \tag{1}$$

$$x_2(n_2) = \mathrm{x}_2\left(\frac{n_2}{(1+\epsilon)f_\mathrm{s}} + \Delta T_{21}\right) \tag{2}$$

Where $\Delta T_{21}$ is the time at which the sampling of $x_2(n_2)$ starts. Here, the sample number that refers to the same time $t$ of channel $i$ $(i = 1, 2)$ is given by

$$n_1 = tf_s, \tag{3}$$

$$n_2 = (1+\epsilon)(t - \Delta T_{21})f_s. \tag{4}$$

Then, $n_2$ is expressed with $n_1$ as below,

$$n_2 = (1+\epsilon)n_1 - (1+\epsilon)D_{21}, \tag{5}$$

where $D_{21} = \Delta T_{21} f_\mathrm{s}$ stands for the discrete time of the first channel when the recording of the second channel starts. And the difference between $n_1$ and $n_2$ is given by

$$\phi(n_1) = n_2 - n_1 = \epsilon n_1 - (1+\epsilon)D_{21}. \tag{6}$$

The difference $\phi(n_1)$ in the number of samples between two signals has a proportionate relationship, and increases with time. This causes the source image to drift and is equivalent to the source position moving artificially. Such movement disrupts conventional microphone array techniques which utilize the time difference of arrival (TDOA) explicitly or implicitly to control directivity. Therefore, it is necessary for the ad-hoc microphone array to estimate the sampling frequency mismatch $\epsilon$.

## III. Supervised identification and compensation of sampling mismatch

In our proposed framework in which short intervals where only a target source is active are given for the speech enhancement, the estimation of the sampling mismatch is no longer an unsupervised estimation problem. This section describes our proposed supervised identification of the sampling mismatch.

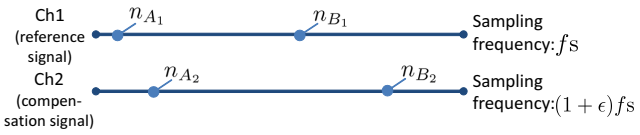


Fig. 1. Mismatch model.

### A. Identification of sampling mismatch using single-source-active short intervals

Before proceeding to a discussion of the estimation procedure, we show that both the sampling frequency mismatch $\epsilon$ and the recording start offset $D_{21}$ are identifiable if two pairs $\{n_{\mathrm{A}1}, n_{\mathrm{A}2}\}$ and $\{n_{\mathrm{B}1}, n_{\mathrm{B}2}\}$ of times corresponding to the same analogue times are available. These variables have to satisfy the following conditions.

$$n_{\mathrm{A}2} = (1+\epsilon)(n_{\mathrm{A}1} - D_{21}), \tag{7}$$

$$n_{\mathrm{B}2} = (1+\epsilon)(n_{\mathrm{B}1} - D_{21}), \tag{8}$$

The conditions identify $\epsilon$ and $D_{21}$ as

$$\epsilon = \frac{n_{\mathrm{B}2} - n_{\mathrm{A}2}}{n_{\mathrm{B}1} - n_{\mathrm{A}1}} - 1, \tag{9}$$

$$D_{21} = \frac{n_{\mathrm{A}1}n_{\mathrm{B}2} - n_{\mathrm{A}2}n_{\mathrm{B}1}}{n_{\mathrm{B}2} - n_{\mathrm{A}2}}. \tag{10}$$

Thus by estimating two pairs of corresponding times $n_{\mathrm{A}i}$ and $n_{\mathrm{B}i}$, $i = 1, 2$, we can obtain the estimate of $\epsilon$ and $D_{21}$. Since precise estimation of these time pairs is difficult and the estimation necessarily has errors, it is preferable that the $n_{\mathrm{A}1}, n_{\mathrm{A}2}$ values are small and those of $n_{\mathrm{B}1}, n_{\mathrm{B}2}$ large.

Now we discuss how to identify a sampling mismatch from a single source activity. We assume that we have two short intervals when one of the sources is active near the beginning and one is active of the end of the recording for each channel, and the estimation is accomplished by analyzing the time difference to maximize the correlation between the channels, and estimate the synchronous time pairs included in the specified intervals, as shown in Fig. 1. However, there are two issues which mean that the estimation cannot be an exact one. The first issue is that the correlation gives only the TDOA of each interval, which affects both the sampling mismatch and the relative positioning of the microphones and the source. Thus it is hard to evaluate only the effect of the sampling mismatch. However, the TDOA caused by the positioning is constant when the source does not move, and we ignore its effect. Although the estimation of the sampling frequency mismatch $\epsilon$ is not effected but the recording start offset $D_{21}$ is given a small error. This error is problematic when the direction of arrival (DOA) is explicitly used in array signal processing. However, it is not problematic in our scenario because the maximum SNR beamformer that we use in the speech enhancement stage uses only the source activity information and the DOA information is not required. The second issue is that the specification of only the intervals is not sufficient for the specification of the synchronous times because where exactly in the single source is located in the roughly specified interval by user's hands. Hereafter, we call

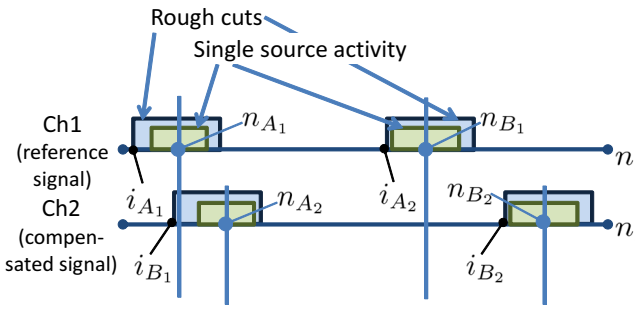


Fig. 2. Two pairs of rough cuts.

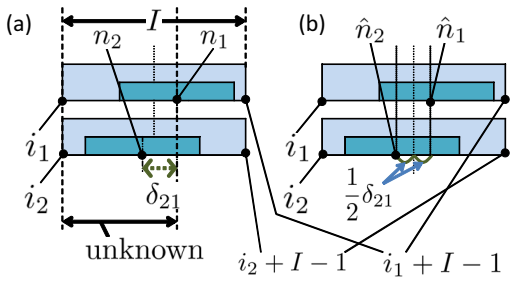


Fig. 3. Correlation in rough cuts.

the roughly specified interval the *rough cut*. We discuss this issue and show how to minimize the $n_1$ and $n_2$ estimation error in the rough cuts in the following.

Suppose we have a pair of rough cuts of the two channels, and the length of both rough cuts is $I$. The rough cuts are denoted by $i_1, \ldots, i_1 + I - 1$ for the first channel and $i_2 + 2, \ldots, i_2 + 2 + I - 1$ for the second channel. As shown in Fig. 2, the time difference of the speech activity in the rough cuts is estimated as

$$\delta_{21} = \arg\max_{\tau} \sum_{l=0}^{I-1} x_1(i_1 + l)\, x_2(i_2 + l - \tau). \tag{11}$$

By ignoring the TDOA caused by the positions as discussed above, the following relation can be assumed.

$$n_2 - n_1 = \delta_{21}. \tag{12}$$

However, the size of $n_1 - i_1$ remains unknown, as shown in Fig. 3 (a). Therefore, as a safe choice of the estimation, we assume that the speech activity is located in the center in average as

$$n_1 + n_2 = i_1 + i_2, \tag{13}$$

and obtain the estimate as

$$n_1 = i_1 + I/2 + \frac{1}{2}\delta_{21}, \tag{14}$$

$$n_2 = i_2 + I/2 - \frac{1}{2}\delta_{21}, \tag{15}$$

as shown in Fig. 3 (b). Although the error remains in the estimate, its effect is reduced by making $n_{\mathrm{A1}}$ and $n_{\mathrm{A2}}$ small, and $n_{\mathrm{B1}}$ and $n_{\mathrm{B2}}$ large.

### *B. Modeling sampling frequency mismatch in short-time frames*

Before we proceed to the STFT analysis, we discuss the effect of drift in a short-time frame. We show that the sampling frequency mismatch can be disregarded in a short interval.

The discrete time of the second channel synchronous with the $(n_1 + m)$-th sample of the first channel is given by the relation in (6) as

$$\begin{aligned}\phi_{21}(n_1 + m; \epsilon, D_{21}) &= (1+\epsilon)(n_1 - D_{21}) + (1+\epsilon)m \\ &= \phi_{21}(n_1; \epsilon, D_{21}) + (1+\epsilon)m,\end{aligned} \tag{16}$$

and can be approximated under the condition $|m\epsilon| \ll 1$ as

$$\phi_{21}(n + m; \epsilon, D_{21}) \approx \phi_{21}(n; \epsilon, D_{21}) + m. \tag{17}$$

Thus the discrete times $n_1 + m$ and $n_2 + m$ of the two channels near the synchronous pair $n_1$ and $n_2$ can be regarded as synchronous.

Therefore, a frame analysis $x_i^{\mathrm{fr}}(l, n_i), l = 0, \ldots, L-1$ of the $i$-th channel of length $L$ (throughout this paper we assume $L$ is even) centered at $n_i$, given by

$$x_i^{\mathrm{fr}}(l, n_i) = w(l)\, x_i\left(l + n_i - \frac{L}{2}\right), \tag{18}$$

where $w(l)$ is an appropriate window function, is almost synchronous between the channels $i = 1, 2$. Since the sampling frequency mismatch $\epsilon$ is generally in the order of $10^{-5}$ and the typical frame length for microphone array signal processing is in the order of 0.1 second, the largest approximation error $|\epsilon L|/2$ of the time in such a frame analysis is usually in the order of 1 µs. Since such the worst error appears near the beginning and the end of the frame, the influence of the errors is reduced by choosing a typical window function $w(l)$ to suppress the amplitude near both ends.

### *C. Synchronization by frame analysis of non-integer sample shift*

Here we discuss the STFT expression of the approximation of $x_2^{\mathrm{fr}}(l, n_2)$ assuming $\epsilon$ and $D_{21}$ are given. The STFT analysis of the $i$-th channel of the frame centered at the sample $n$ is given by

$$X_i(k, n) = \sum_{l=0}^{L-1} x_i^{\mathrm{fr}}(l, n) \exp\left(-\frac{2\pi \jmath k l}{L}\right), \tag{19}$$

where $k = -L/2, \ldots, L/2 - 1$ is the discrete frequency index. Note that the transform is calculated by using a fast Fourier transform in practical processing. According to (6), the discrete time of the second channel synchronous to the central time $n_1$ of $X_1(k, n_1)$ is given by $n_2 = \phi_{21}(n_1; \epsilon, D_{21})$. In [6], we approximated the STFT of the second channel centered at the non-integer time with the following equation.

$$\begin{aligned}&\tilde{X}_2(k, \phi_{21}(n_1; \epsilon, D_{21})) = \\ &X_2(k, n_1) \exp\left(\frac{2\pi \jmath k (\phi_{21}(n_1; \epsilon, D_{21}) - n_1)}{L}\right).\end{aligned} \tag{20}$$

However, this linear phase compensation assumes that the size $|\phi_{21}(n_1;\epsilon,D_{21})-n_1|$ of the shift is much smaller than the frame size $L$, which cannot be maintained with long observation. To avoid the error that arises with the mismatch of the assumption, we apply the frame analysis with the nearest integer central time, and compensate for the effect of the rounding by the circular time shift using a linear phase filter.

The integer sample $\bar{\phi}_{21}(n_1;\epsilon,D_{21})$ nearest to the desired central time $\phi_{21}(n_1;\epsilon,D_{21})$ is given by

$$\bar{\phi}_{21}(n_1;\epsilon,D_{21}) = \arg\min_{n\in\mathbb{Z}} |\phi_{21}(n_1;\epsilon,D_{21})-n|. \quad (21)$$

Since the central sample $\bar{\phi}_{21}(n_1;\epsilon,D_{21})$ of the short-time frame $x_2^{\mathrm{fr}}\left(l,\bar{\phi}_{21}(n_1;\epsilon,D_{21})\right)$ is delayed from the non-integer time $\phi_{21}(n_1;\epsilon,D_{21})$ by $\tilde{\phi}_{21}(n_1;\epsilon,D_{21})$, given by

$$\tilde{\phi}_{21}(n_1;\epsilon,D_{21}) = \phi_{21}(n_1;\epsilon,D_{21}) - \bar{\phi}_{21}(n_1;\epsilon,D_{21}), \quad (22)$$

we obtain the approximation of synchronization in the STFT domain by compensating for the delay with the linear phase filter as

$$\hat{X}_2(k,\phi_{21}(n_1;\epsilon,D_{21})) = X_2\left(k,\bar{\phi}_{21}(n_1;\epsilon,D_{21})\right)\exp\left(\frac{2\pi\jmath k\tilde{\phi}_{21}(n_1;\epsilon,D_{21})}{L}\right). \quad (23)$$

To obtain the STFT analysis for array signal processing, the central samples $n_1$ of the first channel should be defined with a regular frame shift, and the second channel has to be adjusted. First, we analyze the first channel as $X_1(k,n_1)$, $n_1 = rR$, $r = 0,1,2,\ldots$, where $R$ is the frame shift appropriate for the signal reconstruction determined by overlap-and-add analysis, and $r$ is the frame index. Second, we obtain the STFT analysis of the second channel as $\hat{X}_2(k,\phi_{21}(n_1;\epsilon,D_{21}))$ with (23). This STFT of the second channel corresponds to a frame analysis with a non-integer frame shift $(1+\epsilon)R$. Note that synchronized observed signals can be obtained by an inverse STFT analysis with the frame shift $R$.

## IV. Speech enhancement of asynchronous recording using maximum SNR beamformer

Next, we describe a maximum SNR beamformer for speech enhancement that employs single source activity. With the synchronization described in the previous section, it is possible to control directivity even for asynchronous multichannel recording. And we are able to achieve optimal speech enhancement that maximizes the SNR. In this section, we adapt the maximum SNR beamformer using the single source activity to time corrected signals.

Here, the power ratio $\lambda(\omega)$ is expressed as

$$\lambda(\omega) = \frac{\mathbf{w}(\omega)\mathbf{R}_{\mathrm{T}}(\omega)\mathbf{w}^{\mathrm{H}}(\omega)}{\mathbf{w}(\omega)\mathbf{R}_{\mathrm{I}}(\omega)\mathbf{w}^{\mathrm{H}}(\omega)}. \quad (24)$$

where, $\mathbf{R}_{\mathrm{T}}$ and $\mathbf{R}_{\mathrm{I}}$ are the covariance matrices of the activity of the target signal, which are expressed as

$$\mathbf{R}_{\mathrm{T}}(\omega) = \frac{1}{|\Theta_{\mathrm{T}}|}\sum_{t\in\Theta_{\mathrm{T}}}\mathbf{x}_{\mathrm{T}}(\omega,t)\mathbf{x}_{\mathrm{T}}^{\mathrm{H}}(\omega,t). \quad (25)$$

$$\mathbf{R}_{\mathrm{I}}(\omega) = \frac{1}{|\Theta_{\mathrm{I}}|}\sum_{t\in\Theta_{\mathrm{I}}}\mathbf{x}_{\mathrm{I}}(\omega,t)\mathbf{x}_{\mathrm{I}}^{\mathrm{H}}(\omega,t). \quad (26)$$

Here, $\Theta_{\mathrm{T}}$ and $\Theta_{\mathrm{I}}$ are sets of the time frames of the target signal interval and the non-target interval, respectively. The filter $\mathbf{w}(\omega)$ used to maximize the power ratio $\lambda(\omega)$ is given as an eigenvector corresponding to the maximum eigenvalue of the following generalized eigenvalue problem;

$$\mathbf{w}(\omega)\mathbf{R}_{\mathrm{T}}(\omega) = \lambda(\omega)\mathbf{w}(\omega)\mathbf{R}_{\mathrm{I}}(\omega). \quad (27)$$

Since the maximum SNR beamformer $\mathbf{w}(\omega)$ has a scaling ambiguity, we revise the beamformer as:

$$\mathbf{w}(\omega) \leftarrow b_k(\omega)\mathbf{w}(\omega), \quad (28)$$

where $b_k(\omega)$ is the $k$-th component of $\mathrm{b}(\omega)$ given by

$$\mathbf{b}(\omega) = \frac{\mathbf{w}(\omega)\mathbf{R}_{\mathbf{x}}(\omega)}{\mathbf{w}(\omega)\mathbf{R}_{\mathbf{x}}(\omega)\mathbf{w}^{\mathrm{H}}(\omega)}, \quad (29)$$

$$\mathbf{R}_{\mathbf{x}}(\omega) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}(\omega,t)\mathbf{x}^{\mathrm{H}}(\omega,t). \quad (30)$$

Then enhanced signal $y(\omega,t)$ is obtained as

$$y(\omega,t) = \mathbf{w}^{\mathrm{H}}(\omega)\mathbf{x}(\omega,t). \quad (31)$$

## V. Experiments

### A. Experimental conditions

We evaluate our proposed speech enhancement strategy for a distributed microphone array scenario using real portable recording devices.

The task is to enhance the desired speech in the mixture consisting of target and interfering speakers' voices observed with two stereo recording devices. The voices are played back from different loudspeakers. Since the objective evaluation of ad-hoc microphone array recording is not simple, we recorded the speech in a special manner that we describe later. Since the effect of the drift in the asynchronous recording is considerable even if the sampling frequency mismatch is small, we observed a 30-minite-long signal. By recording the same signals for the training and the evaluation at both the beginning and the end of the 30-minute recording, and we were able to evaluate the following two conditions:

a) Adopt the beamformer at the beginning and apply it to the speech enhancement at the beginning.
b) Adopt the beamformer at the beginning and apply it to the speech enhancement at the end.

Needless to say, condition b) is the hardest. We compare the following three methods.

i) Beamforming two channels of one device. (Fig. 4)
ii) Beamforming signals, where their recording start is roughly aligned using the recordings of the training intervals. (Fig. 5)

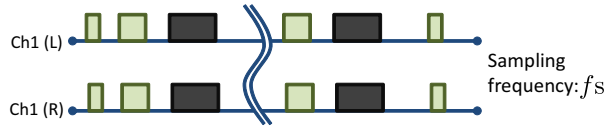


Fig. 4. i) Beamforming of two channels of one device.

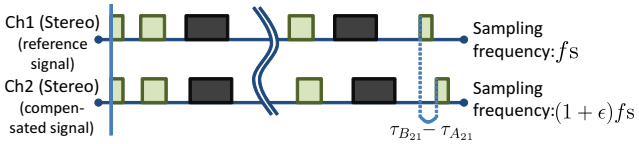


Fig. 5. ii) Beamforming the signals with their recording start is roughly aligned using the recordings of the training intervals.

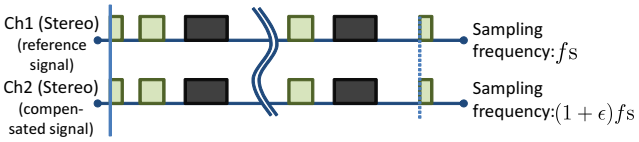


Fig. 6. iii) Beamforming the signals of the compensated sampling (proposed method).

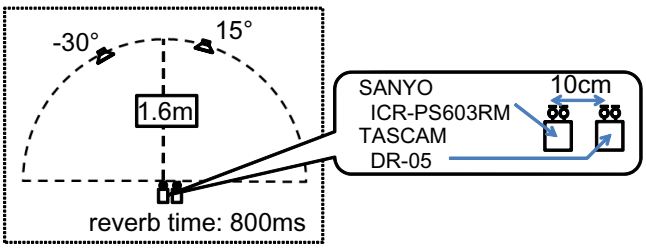


Fig. 7. Recording room.

iii) Beamforming the signals of the compensated sampling (proposed method). (Fig. 6)

We show the recording room layout in Fig.7. The reverberation time T60 is 800 ms. We played back female and male speakers' voices from two loudspeakers, and we regarded the female boice as the desired speech. The recording devices we used were SANYO ICR-PS603RM and TASCAM DR-05. Both of the devices have the nominal sampling frequency of 16,000 Hz and the quantization resolution of 16 bit. The lengths of the signals used for training the desired signal and the interference were 5 ms, and the mixed signal for the evaluation was 30 ms. The frame length and the frame shift were consisted of 16,384 samples and 8,192 samples, respectively. The evaluation scores were the signal-to-distortion ratio (SDR) as a quality measure and the signal-to-interference ratio (SIR) as the interference reduction score.

To obtain an ideally separated signal as the reference for the objective evaluation, we observed each of the sources separately with the same layout of the microphones and the loudspeakers, and we synthesized the observed mixture

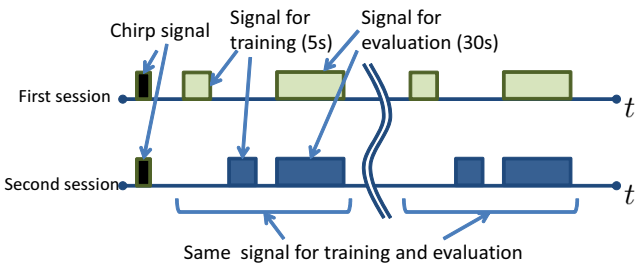


Fig. 8. Recorded signal of each session.

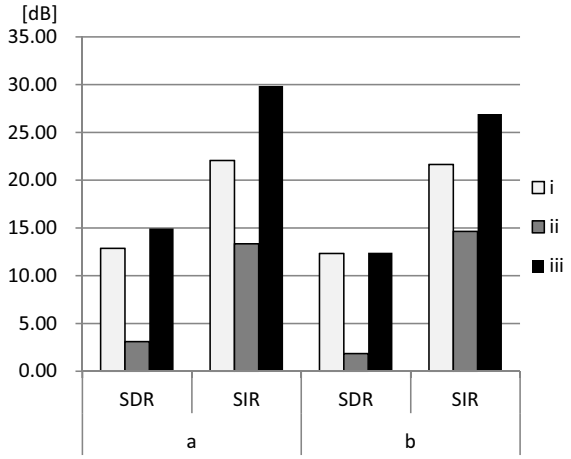


Fig. 9. Experimental results.

by summing those two observations. However, we have to consider the different asynchronous conditions of these two observation sessions; the two asynchronous devices start recording at different times, and the same difference cannot be reproduced again. Thus we align the recording start time of the second observation with that of the first by playing a chirp signal at the beginning of both sessions. By giving a shift to the observation of the second session to maximize the correlation of the two observed chirps, the starts of the recording of the two sessions are aligned. Note that we assume that the sampling frequency of a device remains unchanged throughout the two recording sessions. We show the structure of the signals played back in each session. (Fig. 8)

### *B. Discussion*

The experimental results are shown in Fig. 9. According to the proposed estimation, the sampling frequency mismatch is about 104 ppm.

The graph shows that recording start time offset compensation alone given as the method ii is insufficient for time synchronization. Thus it can be said that the effect of the drift is severe enough to degrade the array signal processing in this situation, and we must compensate for the sampling frequency mismatch. Needless to say, method i using the two synchronized channels is not affected by the drift, and it performs the speech enhancement successfully throughout the

recording. The proposed method given as method iii performs better than method i. Therefore we can conclude that the proposed method successfully compensates for the sampling mismatch and utilizes the asynchronous channels effectively for speech enhancement.

## VI. Conclusion

In this paper, we proposed a speech enhancement framework based on an ad-hoc microphone array using single source activity. The single source activity is utilized both in the synchronization stage and the subsequent array signal processing stage. Experimental results showed that the proposed method effectively uses the asynchronous recording channels with drift for speech enhancement.

## Acknowledgment

This project received the support of the National Institute of Informatics (NII) as part of their promotion of strategic research named "Grand Challenge".

## References

[1] O. L. Frost. "An algorithm for linearly constrained adaptive array processing." Proc. IEEE, Vol. 60, No. 8, pp. 926–935, August 1972.

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

[3] 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, Sep. 2010.

[4] Enrique Robledo-Arnuncio, Ted S. Wada, and Biing-Hwang Juang, "On dealing with sampling rate mismatches in blind source separation and acoustic echo cancellation," Proc. WASPAA, pp.21-24, 2007

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

[6] S. Miyabe, N. Ono, and S. Makino, "Blind compensation of inter-channel sampling frequency mismatch with maximum likelihood estimation in STFT domain," *Proc. ICASSP,* pp..674-678, 2013.

[7] N. Ono, H. Kohno, N. Ito, and S. Sagayama, "Blind Alignment of Asynchronously Recorded Signals for Distributed Microphone Array," Proc. WASPAA, pp.161-164, Oct., 2009.

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

[9] H. L. Van Trees, ed., Optimum Array Processing, Wiley, 2002.

[10] 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.

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