Acoustic Object Canceller Using Blind Compensation for Sampling Frequency Mismatch

Takao Kawamura[1], Nobutaka Ono[2], Robin Scheibler[2], Yukoh Wakabayashi[2], Ryoichi Miyazaki[1]
[1] *National Institute of Technology, Tokuyama College, Yamaguchi, Japan*
{i14kawamura,miyazaki}@tokuyama.kosen-ac.jp
[2] *Tokyo Metropolitan University, Hino, Tokyo, Japan*
{onono,robin,wakayuko}@tmu.ac.jp

***Abstract*—In this paper, we propose a method of removing a known interference from a monaural recording. Generally, the elimination of a nonstationary interference from a monaural recording is difficult. However, if it is a known sound, such as the ringtone of a cell phone, radio and TV broadcasts, and commercially available music provided by a CD or streaming, their signals can be easily obtained. In our proposed method, we define such interference as an *acoustic object*. Although the sampling frequencies of the recording and the available acoustic object might be mismatched, we compensate the mismatch and remove the acoustic object from the recording by maximum likelihood estimation using the auxiliary function technique. We confirm the effectiveness of our method by experimental evaluations.**

***Index Terms*—noise suppression, noise canceller, acoustic object, sampling frequency mismatch, auxiliary function**

I. Introduction

Unlike multichannel recording, to which various array signal processing techniques can be applied, it is generally challenging to remove nonstationary noise from a monaural recording. Some algorithms [1]–[3] for noise suppression are based on the estimation of a noise power spectrum, but the accuracy of noise estimation is imperfect. However, if the interference is a known sound, such as the ringtone of a cell phone, radio and TV broadcasts, and commercially available music provided by a CD or streaming, their signals can be easily obtained. In our study, we define such interference as an *acoustic object* and focus on it. We treat it as a new channel and remove it from a monaural recording with high precision by an array signal processing method. However, the sampling frequencies of the recording and the available acoustic object can be mismatched. An asynchronous microphone array [4]–[8], which consists of independent recording devices, also has a sampling frequency mismatch. Such a mismatch degrades the performance of signal processing [4], [9], [10]. Thus, a method of compensating for the sampling frequency mismatch has been proposed [4], [11]–[15] for asynchronous microphone arrays. In this study, the monaural recording and the acoustic object are treated as components of an asynchronous microphone array, and we use techniques that compensate for the sampling frequency mismatch [12], [13]. Then, the frequency response of the acoustic object is determined by the maximum likelihood estimation using the auxiliary function method, also known as the majorization minimization (MM) algorithm [16], so the acoustic object is removed from the recording. We confirm the effectiveness of our proposed method by experimental evaluations.

II. Problem Formulation

Suppose a situation that we record sound by a monaural microphone, but a known signal interferes with the recording. A typical situation is recording speech interfered with by the background music. Let $x(t)$ be a recorded signal that is modeled by

$$x(t) = o(t - t_d) * h(t) + s(t), \tag{1}$$

where $o(t), h(t)$, and $s(t)$ are a known signal, the impulse response from the source of $o(t)$ to the microphone, and a remaining component, respectively. All the signals are represented in the continuous time domain. The objectives of this study are to remove the contribution of $o(t)$ from $x(t)$ and to estimate $s(t)$ under the assumption that $o(t)$ is known. Hereafter, we refer to $o(t)$ and $s(t)$ as the acoustic object signal and the target signal, respectively. The variable t_d indicates the time difference between $x(t)$ and $o(t)$.

Since the acoustic object signal $o(t)$ and the recorded signal $x(t)$ are sampled by different analog-to-digital converters, there can be a mismatch of the temporal position and the sampling frequencies mainly caused by the individual variability of the quartz in their clock generators. Let $x[n]$ and $o[n]$ be the discrete-time representation of the recorded signal and the acoustic object signal, respectively. We assume that the analog-to-digital converters used to obtain $x[n]$ and $o[n]$ have a common nominal sampling frequency, but the actual sampling frequencies have a small mismatch that is represented as an unknown dimensionless time-invariant scalar ϵ_o ($|\epsilon_o| \ll 1$). Then, $x[n]$ and $o[n]$ can be expressed as

$$x[n] = x\left(\frac{n}{f_x}\right) = x\left(\frac{n}{(1+\epsilon_o)f_o}\right), \tag{2}$$

$$o[n] = o\left(\frac{n}{f_o}\right), \tag{3}$$

where f_x and f_o indicate the sampling frequencies of the recorded and acoustic object signals, respectively.

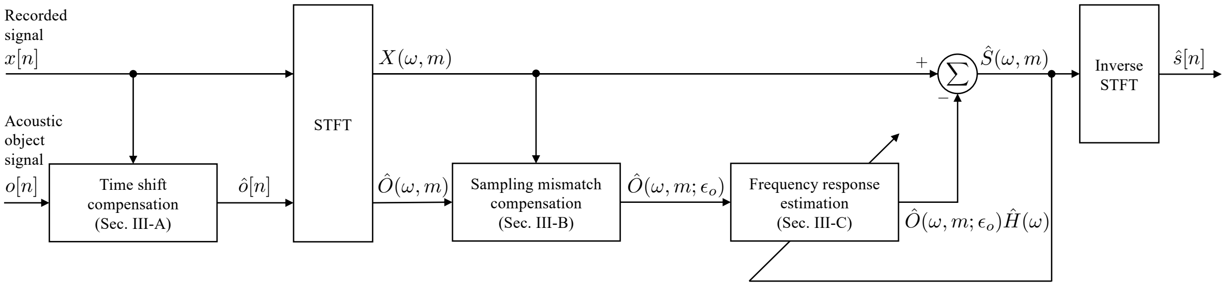
Recorded signal
$x[n]$
Acoustic object signal
$o[n]$
Time shift compensation (Sec. III-A)
$\hat{o}[n]$
STFT
$X(\omega, m)$
$\hat{O}(\omega, m)$
Sampling mismatch compensation (Sec. III-B)
$\hat{O}(\omega, m; \epsilon_o)$
Frequency response estimation (Sec. III-C)
$\hat{O}(\omega, m; \epsilon_o)\hat{H}(\omega)$
+
−
$\hat{S}(\omega, m)$
Inverse STFT
$\hat{s}[n]$

Fig. 1. Overview of procedures in proposed acoustic object canceller.

III. Acoustic Object Canceller

In our study, we propose a framework for removing the acoustic object signal from the recorded signal. This problem is similar to echo canceller and noise canceller in the sense that the interference signal is available. However in our problem, we assume that the recorded signal and the acoustic object signal are recorded by different microphones. In this sense, this is a kind of an asynchronous microphone array.

Generally, an asynchronous microphone array consists of asynchronous multichannel signals recorded by multiple individual recording devices. It has a problem that the performance of array signal processing, including blind source separation, is significantly degraded [4], [9], [10] because the recording in each channel is not synchronized by the difference of the start of the recording and the sampling frequency mismatch. For such a problem, Miyabe *et al.* have proposed a blind synchronization technique of sampling frequencies [12]–[14]. In this research, we apply the above-mentioned blind synchronization technique.

Figure 1 shows an overview of procedures in proposed acoustic object canceller. First, we synchronize between the recorded and acoustic object signals by a rough time shift. Next, the blind compensation for the sampling frequency mismatch is applied to the time-shifted acoustic object signal. Finally, we determine the frequency response of the acoustic object signal by maximum likelihood estimation using the auxiliary function approach and remove the acoustic object signal from the recorded signal using the frequency response.

A. Time shift compensation

The accurate estimation of the time difference t_d in Eq. (1) is not always easy under the sampling frequency mismatch. However, the estimation does not have to be perfectly correct because the small estimation error can be compensated for through the estimation of the frequency response, as will be described in Sec. III-C. We are assuming $|\epsilon_o| \ll 1$, so $x[n]$ and $o[n]$ have a sufficiently high correlation without the compensation for the sampling frequency mismatch. Thus, the discrete time difference τ is estimated simply by maximizing the cross-correlation between $x[n]$ and $o[n]$ as

$$\hat{\tau} = \underset{\tau}{\operatorname{argmax}} \left\{ \sum_n o[n-\tau]x[n] \right\}. \tag{4}$$

Then, we define the time-shifted version of $o[n]$ as $\hat{o}[n] = o[n-\hat{\tau}]$.

B. Sampling mismatch compensation

The sampling frequency mismatch compensation technique has been proposed in [12], [13]. By using this technique, we synchronize the sampling frequencies of the monaural recording $x[n]$ and the time-shifted acoustic object signal $\hat{o}[n]$. We use the same assumptions and approximations as those in the application of sampling mismatch compensation in [17], [18]. On the basis of the assumptions that the sources do not move and are stationary, and the approximation that the time-varying time difference between channels caused by the sampling frequency mismatch is constant within a time frame, the sampling frequency mismatch ϵ_o is compensated for by a linear phase shift in the short-time Fourier transform (STFT) domain. The sampling frequency mismatch ϵ_o is estimated by maximizing the likelihood of the model where the compensated STFT representations follow the time-invariant multivariate Gaussian distribution. The estimation of the sampling frequency mismatch can be iteratively performed to improve the accuracy. The signal processing is detailed in [12], [13].

C. Estimation of frequency response via auxiliary function technique

From Eq. (1), when the length of the impulse response $h(t)$ is sufficiently smaller than the frame length of STFT, the target signal can be estimated in the STFT domain as

$$S(\omega, m) = X(\omega, m) - \hat{O}(\omega, m; \epsilon_o)H(\omega), \tag{5}$$

where $H(\omega)$ is the frequency response of the acoustic object signal and $\hat{O}(\omega, m; \epsilon_o)$ is the STFT representation of the acoustic object signal after the sampling frequency mismatch is compensated as described in the previous subsection. Since $H(\omega)$ is the only unknown factor in Eq. (5), we focus on how to estimate it.

In this study, we assume that $S(\omega, m)$ follows the zero-mean symmetric complex generalized normal distribution, the frequency response $H(\omega)$ is time-invariant, and $S(\omega, m)$ and $\hat{O}(\omega, m; \epsilon_o)$ are uncorrelated. The probability density function of the zero-mean symmetric complex generalized normal distribution is shown as

$$f(x) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-(|x|/\alpha)^{\beta}}, \tag{6}$$

where α and β are the scaling and shape parameters, respectively. It includes a complex normal distribution when $\beta = 2$ and a complex Laplace distribution when $\beta = 1$. Hereafter, we consider $0 < \beta \leq 2$ that corresponds to super Gaussian distribution.

Based on these assumptions, $H(\omega)$ is estimated by maximizing the following log likelihood function:

$$\log L(H(\omega)) = -\frac{1}{\alpha^\beta}\sum_m |S(\omega, m)|^\beta + \mathrm{Const.}, \quad (7)$$

where Const. is a parameter-independent constant. Note that, in the case of $\beta = 2$, maximizing Eq. (7) is equivalent to minimizing the power of the residual signal $S(\omega, m)$, which has been commonly used in the conventional echo canceller and noise canceller. Since the optimization problem to maximize Eq. (7) in terms of $H(\omega)$ has no closed-form solutions in the case of $\beta \neq 2$, we solve it iteratively by applying the auxiliary function method.

In the auxiliary function method, it is necessary to find an appropriate auxiliary function for the objective function. According to the theorem described in [19], for the continuous and differentiable even function $G(x)$ of x, if $G'(x)/x$ is continuous, $x > 0$, positive, and monotonically decreasing,

$$G(x) \leq \frac{G'(x_0)}{2x_0}x^2 + \left(G(x_0) - \frac{x_0 G'(x_0)}{2}\right) \quad (8)$$

holds for any x, and the equality condition is $x = \pm x_0$. The function of x^β appeared in the first term of Eq. (7) satisfies the condition of $G(x)$ when $0 < \beta \leq 2$; thus, the auxiliary function $Q(H(\omega), H_0(\omega))$ can be derived as

$$Q(H(\omega), H_0(\omega)) = \frac{\beta|S_0(\omega, m)|^{\beta-2}}{2}|S(\omega, m)|^2 + \mathrm{Const.}, \quad (9)$$

where $S_0(\omega, m) = X(\omega, m) - \hat{O}(\omega, m; \epsilon_o)H_0(\omega)$ and $H_0(\omega)$ denotes an auxiliary variable. Equation (9) is a quadratic function of $H(\omega)$ since $S(\omega, m)$ includes $H(\omega)$; thus, we can minimize it by differentiating about $H(\omega)$ and setting $H_0(\omega)$ to $H(\omega)^{(k)}$. The update formulae of $S(\omega, m)$ and $H(\omega)$ are given by

$$\hat{S}(\omega, m)^{(k)} = X(\omega, m) - \hat{O}(\omega, m; \epsilon_o)\hat{H}(\omega)^{(k)}, \quad (10)$$

$$\hat{H}(\omega)^{(k+1)} = \frac{\sum_m \hat{O}^*(\omega, m; \epsilon_o)X(\omega, m)/|\hat{S}(\omega, m)^{(k)}|^{\beta-2}}{\sum_m |\hat{O}(\omega, m; \epsilon_o)|^2/|\hat{S}(\omega, m)^{(k)}|^{\beta-2}}, \quad (11)$$

where $\{\cdot\}^*$ indicates the complex conjugate operator. By applying these updates sufficiently, the estimated target signal $\hat{S}(\omega, m)$ is obtained as $\hat{S}(\omega, m)^{(k)}$.

IV. Experimental Evaluations

To confirm the effectiveness of the proposed method, the performance of removing the acoustic object signal was objectively evaluated from the following two perspectives: (i) with/without the compensation for the sampling frequency mismatch and (ii) the change in the shape parameter β.

A. Experimental conditions

We conducted the experiments for both simulated and real-recorded data. For making simulated data, we used Pyroomacoustics [20]. A $4.1 \times 3.8 \times 2.8$ m^3 virtual room was considered and the room absorption was set to 0.2, which corresponds to T_{60} of 0.18 s. Figure 2 shows the arrangement of the loudspeakers and the microphone. The target signal and acoustic object signal were played from loudspeakers (A) and (B), respectively.

As a target signal, we used the speech signal that was made by the concatenation of word utterances chosen from the Japanese Newspaper Article Sentences (JNAS) corpus [21]. The acoustic object signals were the following three types of music: *Solo*, *Ensemble*, and *Chorus*, which are violin solo part of Sonata No. 5 in F major Op. 24, String Quartet No. 14 in G major K. 387, and Chorus (a cappella) of Natsu no omoide composed by Y. Nakada, respectively. They were chosen from SMILE 2004 sound database [22].

For the performance evaluation, we defined the input SNR and output SNR by $10\log_{10}(\sum_n s[n]^2)/(\sum_n(o[n] * h[n])^2)$, $10\log_{10}(\sum_n s[n]^2)/(\sum_n(\hat{s}[n] - s[n])^2)$, respectively, where the summation of n was taken for the time period when either the target signal or the acoustic object signal is not silent. The target signal and the acoustic object signal were mixed at $-5, 0, 5$, and 10 dB of input SNRs. The original sampling frequencies of the recorded signal and acoustic object signal were $16,000$ Hz, and the sampling frequency mismatch was simulated by resampling the recorded signal from $16,000$ to $16,001$ Hz. For STFT, the fast Fourier transform (FFT) was performed at $8,192$ points with $4,096$ length Hamming window, and a shift length was half of the window length. The number of iterations of Eq. (11) was set to 10. To confirm the contribution of the introduced zero-mean symmetric complex generalized normal distribution, the various shape parameter β from 0.2 to 2.0 in steps of 0.2 was used.

In a real-environment experiment, the loudspeakers and a microphone were installed in a laboratory environment of $4.1 \times 3.8 \times 2.8$ m^3 as shown in Fig. 2. We recorded the target signal and the acoustic object signal separately, and synthesized the recorded signal generated by mixing the recorded target signal and recorded acoustic object signal at $-5, 0, 5$, and 10 dB of input SNRs. The sampling frequency of the microphone was $16,000$ Hz, and the other conditions were the same as the previous simulation experiment.

B. Experimental results

We objectively evaluated the effectiveness of the acoustic object signal with output SNR (Figs. 3 and 4). The bar graph corresponds to $\beta = 0.2, 0.4, \cdots, 2.0$ from the left. The white and shaded bars indicate the output SNR obtained from the proposed method with/without the blind synchronization of the sampling frequency mismatch [Sync. (on) and Sync. (off)], respectively. From the results of both experiments, the output SNR was significantly improved by applying the sampling frequency mismatch compensation. While, the output SNR was almost not improved without the sampling frequency

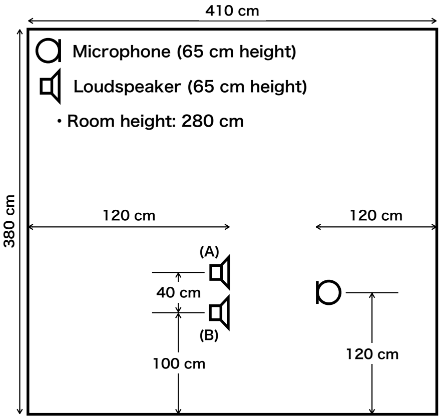

Fig. 2. Locations of loudspeakers and microphone in simulation and recording. Loudspeakers (A) and (B) are for target speech and acoustic object.

mismatch compensation. The reason is that if there is a sampling frequency mismatch, the frequency response will not be apparently time-invariant, and the model expressed by Eq. (5) will not hold. Moreover, it is confirmed that the shape parameter β to obtain the maximum output SNR differs depending on the type of the acoustic object signal and the input SNR. Therefore, using the proposed method is a more flexible model than assuming a complex normal distribution and a complex Laplace distribution; thus, we can obtain a higher output SNR to select the appropriate β. The output SNR in the real environments was slightly lower than that in the simulation. It might be caused by the nonlinearity of the loudspeaker.

Figure 5 shows the examples of spectrograms of (a) the recorded signal generated by adding *Solo* with 10 dB in the simulation, (b) the target signal, and (c) and (d) estimated target signals with/without blind compensation sampling frequency mismatch. The shape parameter was set to 0.8, which provided the highest output SNR in the previous experiment. By focusing on the area surrounded by broken lines in Fig. 5, we can see that the acoustic object signal (harmonic components of the violin) was almost removed by the proposed method with the blind compensation for the sampling frequency mismatch (Fig. 5(c)) compared with the one without the compensation (Fig. 5(d)).

V. Conclusion

In this study, we proposed a method of removing the acoustic object signal from the recorded signal containing the acoustic object signal with high accuracy. In the experiments, we confirmed the output SNR to be high using various sounds in the simulation and real environments. In the future, we plan to simultaneously remove multiple acoustic object signals when those signals are played. In addition, we plan to conduct experiments considering the case where the target signal is music or from multiple speakers. In this study, we did not consider the effect of the nonlinearity of the loudspeaker. We will take it into account in future work.

Acknowledgment

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Numbers JP16H01735 and JP17H00749.

References

[1] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *IEEE Transactions on Speech Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[2] R. C. Hendriks *et al.*, “MMSE based noise PSD tracking with low complexity,” *Proc. ICASSP*, pp. 4266–4269, 2010.

[3] T. Gerkmann and R. C. Hendriks, “Unbiased MMSE-based noise power estimation with low complexity and low tracking delay,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

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

[5] I. Himawan *et al.*, “Clustered blind beamforming from ad-hoc microphone arrays,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 661–676, 2011.

[6] M. Souden *et al.*, “Location feature integration for clustering-based speech separation in distributed microphone arrays,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 354–367, 2014.

[7] D. Cherkassky and S. Gannot, “Blind synchronization in wireless acoustic sensor networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 651–661, 2017.

[8] M. H. Bahari *et al.*, “Blind sampling rate offset estimation for wireless acoustic sensor networks through weighted least-squares coherence drift estimation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 674–686, 2017.

[9] R. Lienhart *et al.*, “On the importance of exact synchronization for distributed audio signal processing,” *Proc. ICASSP*, pp. IV-840–IV-843, 2003.

[10] E. R.-Arnuncio *et al.*, “On dealing with sampling rate mismatches in blind source separation and acoustic echo cancellation,” *Proc. WASPAA*, pp. 34–37, 2007.

[11] S. M.-Golan *et al.*, “Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming,” *Proc. IWAENC*, 2012.

[12] S. Miyabe *et al.*, “Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation,” *Signal Processing*, vol. 107, pp. 185–196, 2015.

[13] S. Miyabe *et al.*, “Blind compensation of inter-channel sampling frequency mismatch with maximum likelihood estimation in STFT domain,” *Proc. ICASSP*, pp. 674–678, 2013.

[14] S. Miyabe *et al.*, “Optimizing frame analysis with non-integrer shift for sampling mismatch compensation of long recording,” *Proc. WASPAA*, 2013.

[15] R. Sakanashi *et al.*, “Speech enhancement with ad-hoc microphone array using single source activity,” *Proc. APSIPA*, 2013.

[16] D. R. Hunter and and K. Lange, “A Tutorial on MM Algorithms,” *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[17] K. Ochi *et al.*, “Multi-talker Speech Recognition Based on Blind Source Separation with Ad hoc Microphone Array Using Smartphones and Cloud Storage,” *Proc. Interspeech*, pp. 3369–3373, 2016.

[18] S. Araki *et al.*, “Meeting recognition with asynchronous distributed microphone array,” *Proc. ASRU*, 2017.

[19] N. Ono and S. Miyabe, “Auxiliary-function-based independent component analysis for super-gaussian sources,” *Proc. LVA/ICA*, pp. 165–172, 2010.

[20] R. Scheibler *et al.*, “Pyroomacoustics: a python package for audio room simulation and array processing algorithms,” *Proc. ICASSP*, pp. 351–355, 2018.

[21] K. Ito *et al.*, “JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research,” *The Journal of Acoustical Society of Japan*, vol. 20, pp. 196–206, 1999.

[22] K. Kawai *et al.*, “Introduction of sound material in living environment 2004 (SMILE 2004): A sound source database for educational and practical purposes,” *The Journal of the Acoustical Society of America*, vol. 120, no. 5, p. 3070, 2006.

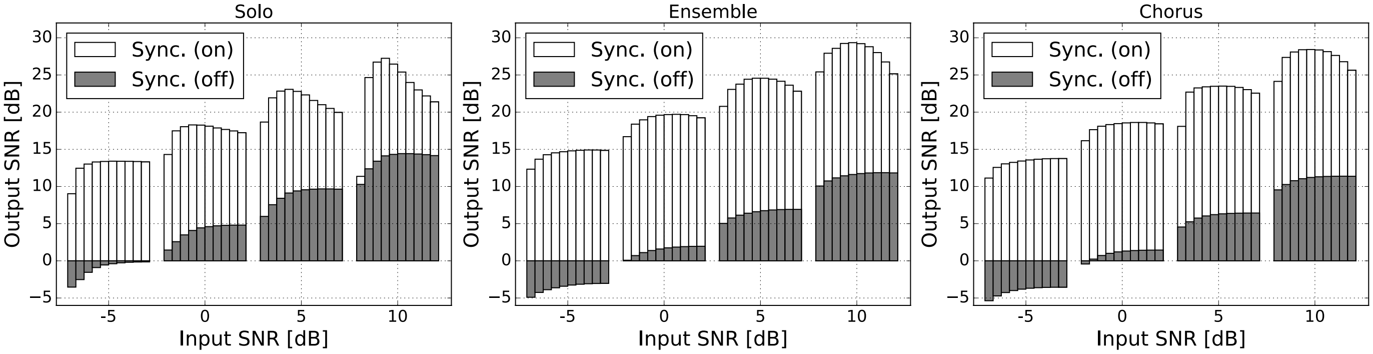

Fig. 3. Output SNR with simulated data using different β values. β increases from 0.2 to 2.0 in steps of 0.2 from left to right.

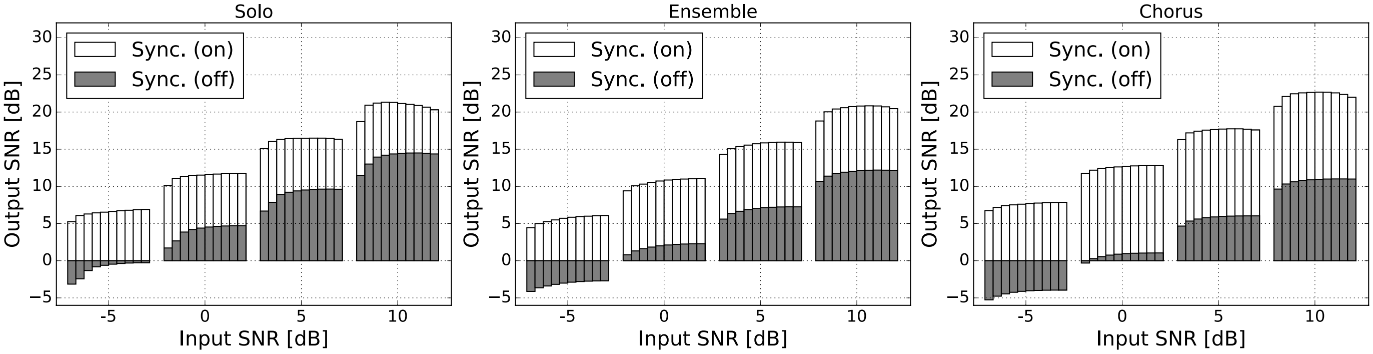

Fig. 4. Output SNR with recorded data using different β values. β increases from 0.2 to 2.0 in steps of 0.2 from left to right.

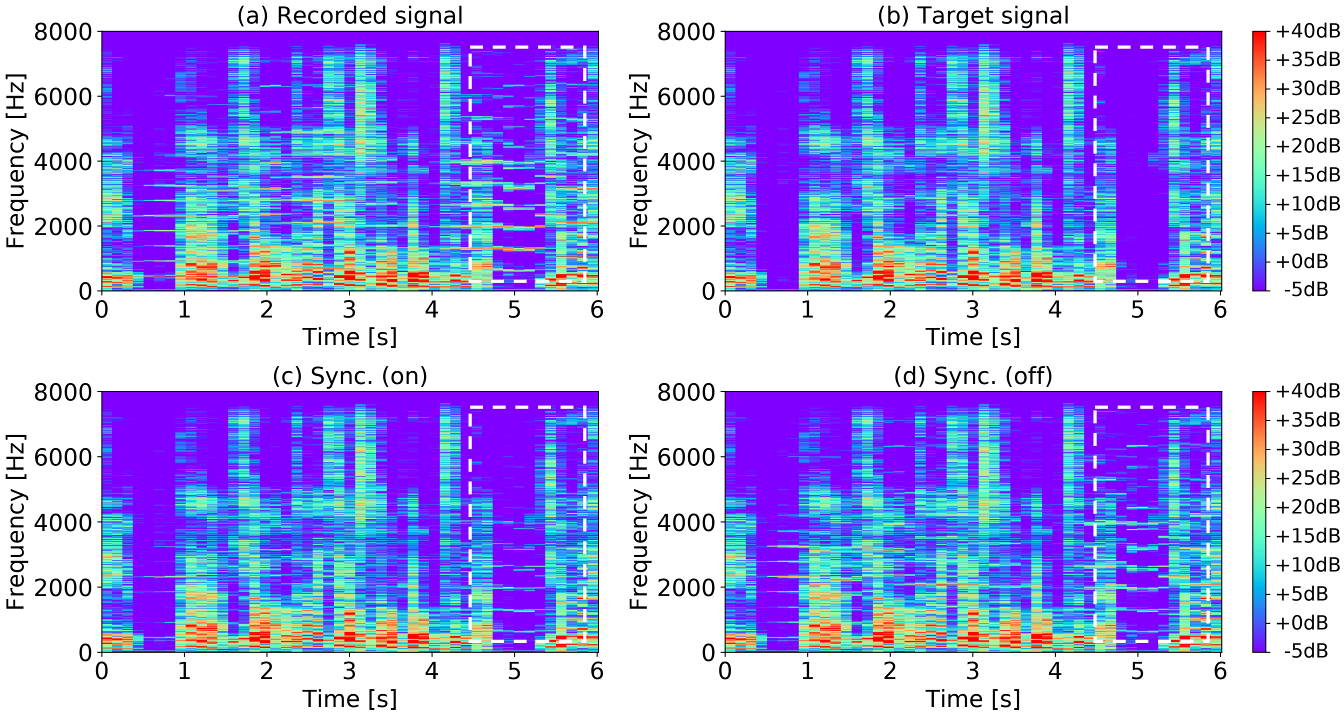

Fig. 5. Spectrograms of (a) recorded signal (target signal was generated by adding *Solo* with 10 dB SNR), (b) target signal, and (c) and (d) estimated target signal with/without blind compensation sampling frequency mismatch, respectively.