Dynamic synchronous averaging for enhancement of periodic signal under sampling frequency variation

Kyosuke Sumiyoshi, Yukoh Wakabayashi, Nobutaka Ono
* Tokyo Metropolitan University, Japan
E-mail: {sumiyoshi-kyosuke@ed., wakayuko@, onono@}tmu.ac.jp

***Abstract*—In this paper, we present a novel method of estimating a room impulse response (RIR) in noisy environments by playing a known periodic signal and recording it for a long time. As is well known, a periodic signal can be easily enhanced, even in a noisy environment, by synchronous averaging. However, in a long-time recording, the sampling frequency of the recording device might fluctuate temporally, which leads to a synchronization error in averaging and degrades the performance of enhancement. To solve this problem, we estimate the time shift between the played-back periodic waveform and the observed signal by their cross-correlation, period by period, and apply synchronous averaging while compensating for the shift dynamically. We also introduce an iterative approach in dynamic synchronous averaging to further improve the performance. In simulation experiments, we confirm that the proposed method effectively enhances the signal and contributes to RIR estimation with high accuracy.**

I. Introduction

A room impulse response (RIR) is an important factor in determining the acoustic characteristics in a room. Generally, high-energy signals, such as a time-stretched pulse [1] and a maximum length sequence [2], and a white noise signal with a high signal-to-noise ratio (SNR) are used to measure an RIR. When we measure an RIR, for example, in a concert hall, it would be desirable to consider the presence of the audience in the hall. It is, however, difficult to measure it in such a situation because a high-energy signal can be annoying to the audience. On the other hand, a low-energy signal cannot be perceived. But, to estimate RIR accurately, it is necessary to enhance the signal because of the low SNR.

Synchronous averaging along the time direction is commonly used for such enhancement. The longer the averaging time is, i.e., the greater the number of additions, the greater the enhancement is. However, we need to consider the fact that synchronous averaging requires accurate synchronization and that inaccurate synchronization degrades a signal. There are some obstacles to accurate synchronization. In this study, we handle the following two problems: the effect of discretization in digital signal processing and the sampling cycle variation of recording devices. Generally, we discretize a continuous-time signal and treat the discrete sample signal. A discrete sample signal is easy to use in a computer, while the discretization of time information complicates synchronization because a gap between the samples is often necessary, that is, subsample (non-integer sample) information is needed for accurate synchronization. Non-integer sample estimation based on the correlation between two signals has been proposed [3, 4]. Regarding the problem of sampling cycle variation, it is well known that the sampling frequency of recording devices changes temporally from a few ppm (parts per million, 10^{-6}) to many hundreds of ppm [5, 6]. This problem, the so-called "mismatch of sampling frequency", has been examined in previous studies [7–10]. This mismatch is also a problem in various fields such as blind source separation and speech enhancement [11, 12]. This small fluctuation markedly affects the performance of synchronization, especially in the case of long-time recording, and the signal enhancement performance also deteriorates as a result.

In this paper, we propose a signal enhancement method with synchronous averaging considering the two aforementioned problems as an initial investigation for measuring an RIR using weak signals. We assume the following situation in this study: we play a known signal repeatedly with a known time interval, which we call a weak periodic signal, from a loudspeaker, we record it for a long time with a distant single-channel microphone whose sampling cycle may vary with time, we enhance one cycle signal using the recorded signal, and we estimate the RIR signal. For enhancement the signal, we propose dynamic synchronous averaging based on the non-integer sample estimate, that is important to achieve the highest correlation between recorded signal and periodic signals. Using the enhanced signal, we estimate the RIR signal. We evaluate the performance of the proposed method in an experimental simulation.

II. Problem formulation

Let us consider playing a source signal from a loudspeaker, recording it with a distant microphone, and estimating the impulse response from the loudspeaker to the microphone in a noisy environment. Let $s(t)$ and $h(t)$ be the source signal and the impulse response in the continuous-time domain, respectively. Then, the observed signal $y(t)$ is expressed as

$$y(t) = x(t) + v(t), \tag{1}$$

$$x(t) = \int_0^\infty s(t-\tau)h(\tau)d\tau, \tag{2}$$

where $x(t)$ is the source image of $s(t)$ and $v(t)$ is the background noise. If $s(t)$ is a periodic signal with period P_c, then $x(t)$ is also a periodic signal because $x(t+P_c) = x(t)$.

Let T be the nominal sampling cycle, namely, the reciprocal of the nominal sampling frequency of the recording device, and assume that it is a constant. The effect of its mismatch

or temporal variation will be discussed in the next section. Let $P = P_c/T$ be the period in the discrete time domain and assume that P is an integer. Then, (1) and (2) can be expressed in the discrete domain as

$$y[n] = x[n] + v[n], \tag{3}$$

$$x[n] = \sum_{m=0}^{\infty} s[n-m]h[m], \tag{4}$$

where $s[n]$ is the discrete signal obtained by sampling $s(t)$, as $s[n] = s(nT)$, and $x[n]$, $y[n]$, $v[n]$, and $h[n]$ are defined in the same way.

If the impulse response $h[m]$ is shorter than P, (4) can be expressed as

$$X(k) = S(k)H(k), \tag{5}$$

where $X(k)$, $S(k)$, and $H(k)$ are the discrete Fourier transforms (DFTs) of $x[n]$, $s[n]$, and $h[n]$ with length P ($n = 0, \ldots, P-1$), respectively. Then, we have

$$H(k) = \frac{X(k)}{S(k)}. \tag{6}$$

Because $s[n]$ is a known signal, $S(k)$ can be computed. Then, $h(m)$ can be estimated by the inverse DFT of (6).

The problem is that we obtain a noisy signal $y[n]$ rather than $x[n]$ directly. Since $x[n]$ is a periodic signal with period P, we can estimate it by applying synchronous averaging to $y[n]$.

First, we segment the recorded signal $y[n]$ by period P as

$$y_m[n] = y[n + mP], \quad n = 0, \ldots, P-1, \tag{7}$$

where $y_m[n]$ is the segmented signal in the mth frame. Then, the synchronous average can be written as

$$\hat{x}[n] = \frac{1}{M} \sum_{m=1}^{M} y_m[n], \tag{8}$$

where M is the total number of frames. Replacing $X(k)$ in (6) by the DFT of $\hat{x}[n]$ ($n = 0, \ldots, P-1$), we can estimate the impulse response $h[n]$.

III. Effect of Sampling Frequency Variation

First, let us consider what happens if the sampling cycle is slightly different from a nominal cycle. Let $y_1[n]$ and $y_2[n]$ be sampled signals that are obtained by sampling with sampling cycles T and $T+\varepsilon$, respectively, where ε is the sampling cycle mismatch. The relationship between $y_1[n]$, $y_2[n]$, and $y(t)$ can be expressed as

$$y_1[n] = y(nT), \quad y_2[n] = y(n(T+\varepsilon)). \tag{9}$$

The difference between the sampling times nT and $n(T+\varepsilon)$ increases over time, as shown in Fig. 1. This will distort the synchronous averaging.

If the sampling cycle mismatch ε is constant, it can be easily estimated and compensated for because $y_2[n]$ is still a periodic signal. However, in a long recording, the sampling cycle T could vary slightly. Let $T + \varepsilon[n]$ be the sampling cycle from the $(n-1)$th sample to the nth sample, where $\varepsilon[n]$ is the fluctuation of the sampling cycle. Then, the discrete signal $y[n]$ obtained by sampling $y(t)$ can be expressed as

$$y[n] = y(nT + \tau[n]), \quad \tau[n] = \sum_{l=1}^{n} \varepsilon[l], \tag{10}$$

where $\tau[n]$ represents the temporal shift of the nth sample from the nth sampling point when the sampling cycle T does not vary. Even if the variations of sampling cycle $\varepsilon[n]$ are small, they accumulate over a long time and become non-negligible. Therefore, the sampling frequency variation distorts the synchronization of the segmented signals over a period, and then, synchronous averaging cannot enhance the signal, as shown in Fig. 2. This is the problem with the synchronous averaging of a long recording.

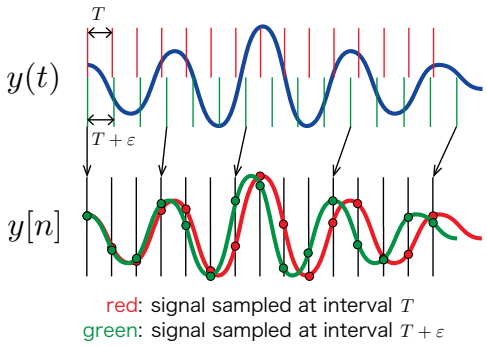

Fig. 1. Conceptual diagram of sampling with different sampling cycles. $y(t)$ and $y[n]$ are continuous and discrete time signals, respectively.

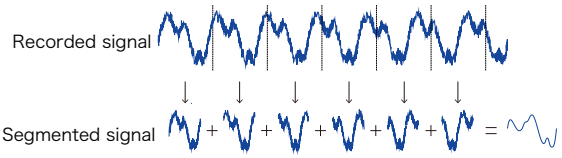

Fig. 2. Difficulty in synchronous averaging under time-varying sampling frequency.

IV. Proposed Method

A. Dynamic synchronous averaging using cross-correlation

In general, synchronous averaging with sampling cycle variation is a difficult task because $\varepsilon[n]$ is unknown. To solve this problem, we propose a dynamic synchronous averaging method with compensation for sampling frequency variation, which consists of four steps, namely, frame segmentation, time delay estimation, synchronous averaging, and their iteration.

First, we segment the recorded signal $y[n]$ by period P of $s[n]$ as in (3). Unlike the discussion in Section II, each frame-by-frame signal waveform does not synchronize owing to expansion and contraction along the time direction caused by sampling frequency variation, as described in Section III. Because the variation of the sampling cycle within one period

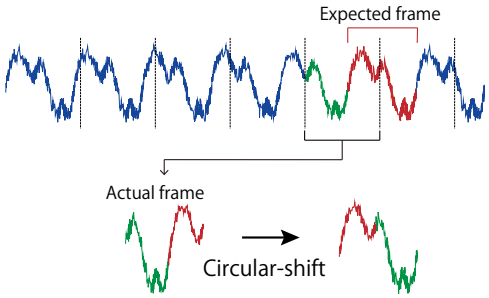

Fig. 3. Conceptual diagram of circular-shift

is very small, it can be considered negligible, and only a circular time shift in a period in $y_m[n]$ is considered. Then, we can estimate the time shift sample τ_m between $s[n]$ and $y_m[n]$ using the cross-correlation between them and its maximum as

$$\tau_m^{(1)} = \arg\max_{\tau} \sum_{n=0}^{P-1} s[n] y_m[n+\tau]. \tag{11}$$

Moreover, we estimate the non-integer sample value of τ to obtain it with non-integer sample accuracy. We perform non-integer sample time delay estimation using parabolic function-based interpolation [13]. This method interpolates the discrete cross-correlation using quadratic functions and estimates a non-integer sample delay using its maximum. The segmented signals are synchronized by the circular-shift for the estimated sample delay $\tau_m^{(1)}$, as shown in Fig. 3. Since the circular-shifted signals $y_m[n+\tau_m^{(1)}]$ are synchronized, we can enhance one period of a signal by synchronous averaging of these signals as

$$\hat{x}^{(1)}[n] = \frac{1}{M} \sum_{m=1}^{M} y_m[n+\tau_m^{(1)}], \tag{12}$$

where M is the total number of frames.

B. Iterative estimation of impulse response and synchronization

The performance of the signal enhancement in (12) depends on the accuracy of the time delay estimation by (11). In a reverberant environment, the convolution of the RIR makes it difficult to estimate the time delay τ because the peaks in the cross-correlation do not occur clearly. Fig. 4 shows the cross-correlation function between the source signal $s[n]$ and the observed signal $y[n]$ when there is no background noise. We can see that there are multiple peaks due to the convolution of the RIR, which complicates finding the correlation between the two signals.

To improve this situation, we consider the re-estimation of the time delay by replacing $s[n]$ in (11) with the enhanced signal $\hat{x}^{(1)}[n]$ in the previous step as follows;

$$\tau_m^{(2)} = \arg\max_{\tau} \sum_{n=0}^{P-1} \hat{x}^{(1)}[n] y_m[n+\tau]. \tag{13}$$

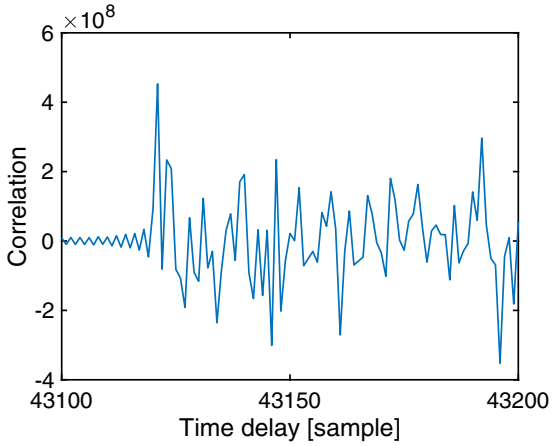

Fig. 4. Cross-correlation function between source and observed signals.

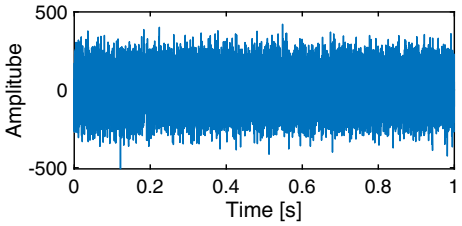

Fig. 5. Source signal $s[n]$.

As described in Section II, we can obtain the signal that is convolved with the RIR from (12). The peaks of the cross-correlation are relatively easy to detect. This iterative approach makes it easy to find the correlation and estimate the time shift. Therefore, we can achieve more accurate synchronous averaging by using $\tau_m^{(2)}$ as follows:

$$\hat{x}^{(2)}[n] = \frac{1}{M} \sum_{m=1}^{M} y_m[n+\tau_m^{(2)}]. \tag{14}$$

To improve the accuracy of the synchronization, it is possible to repeat (13) and (14).

V. Experimental Evaluation

We conducted simulation experiments to verify the efficacy of the proposed method. The experiments consist of two parts. The first part evaluates the enhancement performance of the proposed method without the iterative approach with (13) and (14) (Section V-B). In the second part, we evaluate the performance of signal enhancement and impulse response estimation by using dynamic synchronous averaging with the iterative approach (Section V-C).

A. Experimental conditions

We supposed a real environment and used a reverberation signal, which was convolved with an impulse response generated by the Polack method [14], and we set the reverberation

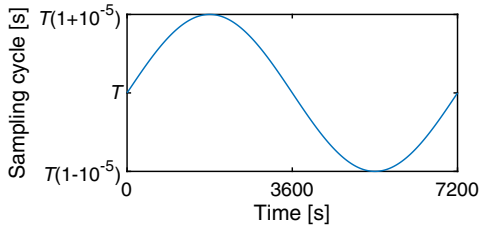

Fig. 6. Time-varying sampling cycle in discrete domain, $T = 1/44100$ [s].

TABLE I
LIST OF FIVE METHODS USED IN COMPARISON METHODS

Name	Time shift	Median filter	Sampling freq. variation
baseline	×	×	not considered
`int sft`	integer	×	considered
`int sft+med`		○	
`non-int sft`	non-integer	×	
`non-int sft+med`		○	

time to about 300 [ms]. The reverberation time has to be shorter than period of source signal P. We used the following periodic signals with a sufficiently wide bandwidth for the source signal $s[n]$ shown in Fig. 5:

$$s[n] = \sum_{k=1}^{K} \cos(2\pi k f n + \phi_k), \tag{15}$$

where K is the number of harmonic components, ϕ_k is the initial random phase of each harmonic component, and f is the fundamental frequency. In this experiment, we set $K = 22,050$ and $f = 1$ [Hz]. We repeated playing of $s[n]$ $7,200$ times while convolving it with the impulse response to create a 2h signal. We set the sampling frequency f_s to $44,100$ [Hz] as the initial value. The sampling cycle $T = 1/f_s$ was varied in the range of $T(1 - 10^{-5})$ to $T(1 + 10^{-5})$, as shown in Fig. 6. By resampling the 2h signal with this variation, we generated the recorded signal $x(t)$ with sampling frequency variation. The period P of the synchronous averaging was $1/fT = 44,100$ samples. We generated noisy speech $y[n]$ by mixing $x[n]$ and an additive white noise signal with a SNR ranging from -60 to 0 [dB]. We used the output SNR as an evaluation measure.

We compared four variants of the proposed method, which are listed in Table I, and a baseline method. As the baseline method, we used synchronous averaging that does not consider the sampling frequency variation. As shown in this table, we employed four types of dynamic synchronous averaging, which include cases in which a median filter is used or not used and cases in which the circular sample shift is conducted with integer precision or non-integer sample precision. We used a median filter under the assumption that the variation of the sampling cycle is small and gradual to reduce the volatility of the estimated τ caused by noise signals. We designed the median filter that the median of the sequence $\tau_{m-50}, \cdots, \tau_{m+49}$ is set to τ_m.

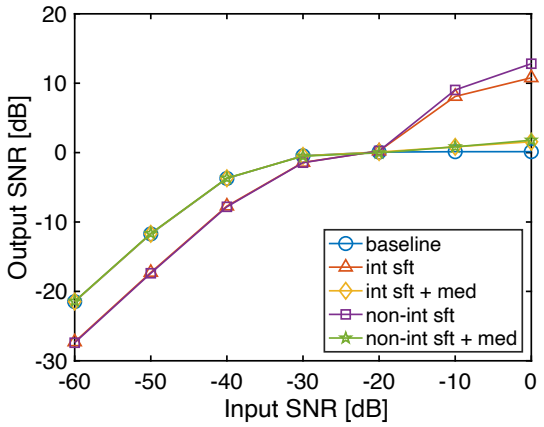

Fig. 7. Relationship between input and output SNRs with five synchronous averaging methods.

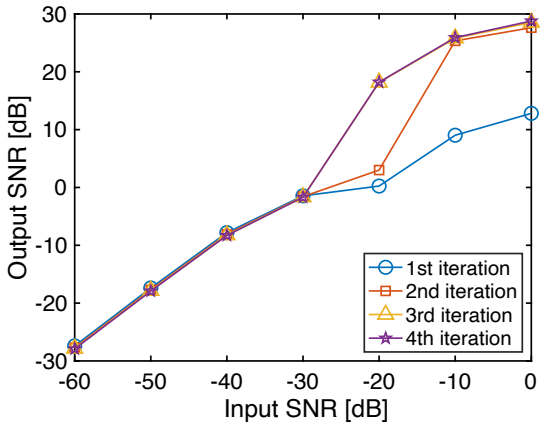

Fig. 8. Relationship between input and output SNRs in each iteration when dynamic synchronous averaging is applied with iterative approach and non-integer sample shift (`non-int sft`).

B. Comparison of signal enhancement with five methods

Fig. 7 illustrates the output SNR with the different methods as a function of the input SNR. Since we add the segmented signals $7,200$ times in all the synchronous averaging methods, the SNR can be expected to improve by $20\log_{10}(\sqrt{7200}) = 38$ [dB]. The tendency of the performance is different for input SNRs of more and less than -20 [dB], and with the cases that the median filter is used and not. At an input SNR of less than -20 [dB], the large estimation errors of the time shift τ when using cross-correlation made accurate synchronous averaging difficult and degraded the enhancement performance, and therefore, the two methods without a median filter, `int sft` and `non-int sft`, were inferior to the baseline method. However, the methods with a median filter, `int sft+med` and `non-int sft+med`, and the baseline method improved SNR by approximately 38 [dB]. Although

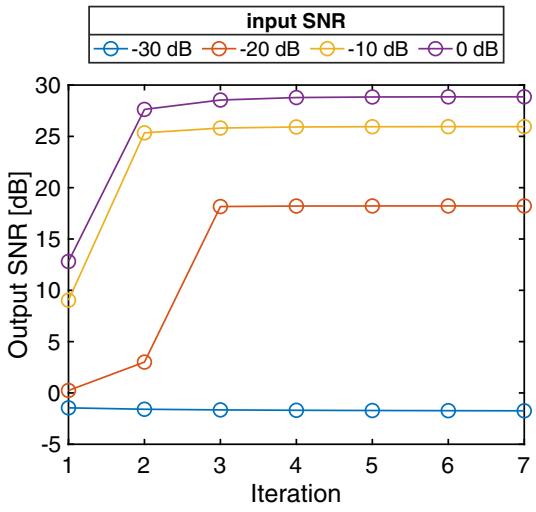

Fig. 9. Change in the enhancement performance at each iteration.

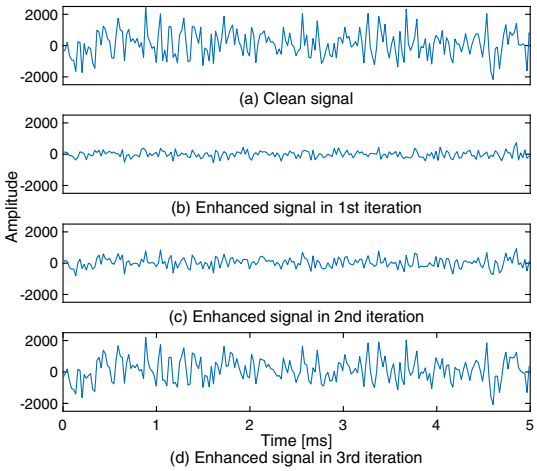

Fig. 10. Change in enhanced signal at 1st, 2nd, and 3rd iterations when input SNR is -20 [dB].

the median filter reduced the errors caused by noise, as we expected, `int sft+med`, `non-int sft+med`, and the baseline method exhibited the same performance. This is because even if the sampling frequency variation is compensated for, the effects of time-shift estimation accuracy on noises are larger than the variation. On the other hand, when the input SNR was more than -20 [dB], all the proposed methods improved the output SNR compared with the baseline method. The output SNRs with `int sft+med` and `non-int sft+med` were lower than those with `int sft` and `non-int sft` because the median filter degraded the estimation accuracy of τ because τ is accurately estimated without a median filter in such higher-SNR environments. Among the methods without a median filter, `non-int sft` outperformed `int sft`. It was confirmed that more accurate synchronization was possible by determining the non-integer sample time shift.

C. Comparison of signal enhancement and RIR estimation by synchronous averaging with the iterative approach

In this section, we evaluate the performance of both signal enhancement and RIR signal estimation with the iterative approach. Here, we used the `non-int sft` method, which achieved the best performance at an input SNR of more than -20 [dB] in Section V-B. The output SNR at each iteration is shown in Figs. 8–9. The result of the first iteration in Fig. 8 is the same as that of `non-int sft` in Fig. 7. As shown, the output SNR improved with each iteration. We confirmed that the accuracy of synchronous averaging is increased by updating the reference signal to estimate the time shift. Also, since Fig. 9 illustrates the performance does not change after the third update, we can see that three updates are sufficient to achieve accurate synchronous averaging. The reason why the enhancement performance does not improve after the third update is that we have achieved marginal performance under the assumption that no variation of the sampling cycle occurs within a frame. Fig. 10 shows the change in enhanced signal at each iteration when the input SNR is -20 [dB]. In this figure, the clean signal (a) is the signal that is obtained by convolving the impulse response and the source signal $s[n]$ with no sampling frequency variation, which is the target signal that we want to obtain by enhancement. We can see that the waveform of the enhanced signal becomes more similar to that of the clean signal with each iteration.

Next, Fig. 11 shows the output SNR of RIR estimation at each iteration. As the performance of synchronous averaging improves, the RIR estimation accuracy also improves. Even when the input SNR was as low as -10 [dB], it was confirmed that we can estimate the RIR with an accuracy exceeding 20 [dB] in the output SNR. Fig. 12 shows the change in estimated RIR at each iteration when the input SNR is -20 [dB]. We can also see that the waveform of the estimated signal is approaches the actual impulse response with each iteration.

VI. Conclusion

In this paper, we proposed a dynamic synchronous averaging method for impulse response estimation in the situation of time-varying sampling frequency. We used cross-correlation and a frame-by-frame circular shift to dynamically detect the time shift caused by the variation of the sampling frequency. We achieved a more accurate synchronization by updating the reference signal to estimate the time shift. Experimental results showed that the proposed method can enhance the target signal even with a low SNR. It was also confirmed that the proposed method is useful for impulse response estimation. In our future work, we will verify the performance of the proposed method in a real environment.

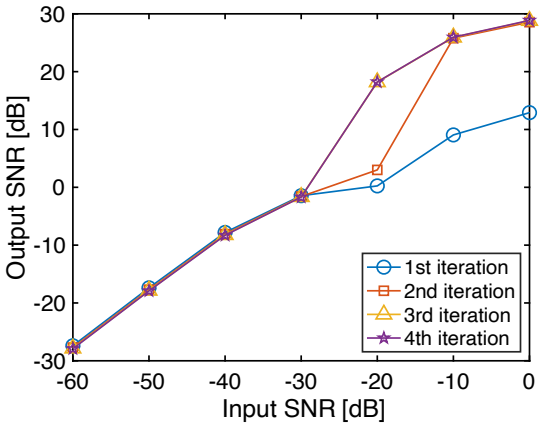

Fig. 11. Output SNR of RIR estimation in each iteration when dynamic synchronous averaging is applied with iterative approach and non-integer sample shift (`non-int sft`).

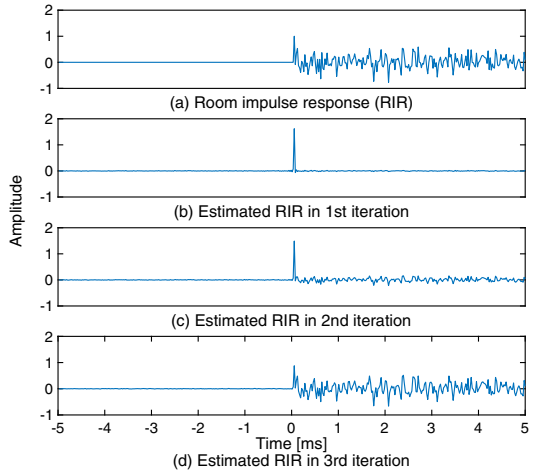

Fig. 12. Change in estimated impulse response at 1st, 2nd, and 3rd iterations when the input SNR is -20 [dB].

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP20H00613.

References

[1] A. J. Berkhout, D. de Vries, and M. M. Boone, "A new method to acquire impulse responses in concert halls," *The Journal of the Acoustical Society of America*, vol. 68, no. 1, pp. 179–183, 1980.

[2] M. R. Schroeder, "Integrated-impulse method measuring sound decay without using impulses," *The Journal of the Acoustical Society of America*, vol. 66, no. 2, pp. 497–500, 1979.

[3] K. Yamaoka, R. Scheibler, N. Ono, and Y. Wakabayashi, "Sub-sample time delay estimation via auxiliary-function-based iterative updates," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 130–134.

[4] L. Zhang and X. Wu, "On the application of cross correlation function to subsample discrete time delay estimation," *Digital Signal Processing*, vol. 16, no. 6, pp. 682–694, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1051200406001230

[5] N. Akira and K. Nobuo, "Various aspects and factors of sampling jitter observed in digital audio products," *Journal of Tokyo University of Information Sciences*, vol. 7, no. 2, pp. 79–92, Feb 2004. [Online]. Available: https://ci.nii.ac.jp/naid/120005455636/

[6] R. Lienhart, I. Kozintsev, S. Wehr, and M. Yeung, "On the importance of exact synchronization for distributed audio signal processing," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings (ICASSP '03)*, vol. 4, 2003, pp. IV–840.

[7] S. Miyabe, N. Ono, and S. Makino, "Estimating correlation coefficient between two complex signals without phase observation," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 421–428.

[8] L. Wang and S. Doclo, "Correlation maximization-based sampling rate offset estimation for distributed microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 571–582, March 2016.

[9] R. Sakanashi, N. Ono, S. Miyabe, T. Yamada, and S. Makino, "Speech enhancement with ad-hoc microphone array using single source activity," in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2013, pp. 1–6.

[10] S. Araki, N. Ono, K. Kinoshita, and M. Delcroix, "Estimation of sampling frequency mismatch between distributed asynchronous microphones under existence of source movements with stationary time periods detection," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 785–789.

[11] E. Robledo-Arnuncio, T. S. Wada, and B. Juang, "On dealing with sampling rate mismatches in blind source separation and acoustic echo cancellation," in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007, pp. 34–37.

[12] H. Chiba, N. Ono, S. Miyabe, Y. Takahashi, T. Yamada, and S. Makino, "Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording," in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 203–207.

[13] G. Jacovitti and G. Scarano, "Discrete time techniques for time delay estimation," *IEEE Transactions on Signal Processing*, vol. 41, no. 2, pp. 525–533, 1993.

[14] J. Polack, "La transmission de l'énergie sonore dans les salles," Ph.D. dissertation, Université du Maine, 1988.