OPTIMIZING FRAME ANALYSIS WITH NON-INTEGRER SHIFT FOR SAMPLING MISMATCH COMPENSATION OF LONG RECORDING

Shigeki Miyabe[1] *Nobutaka Ono*[2] *Shoji Makino*[1]

[1] University of Tsukuba, Life Science Center of Tsukuba Advanced Research Alliance
1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8573, Japan, {miyabe, maki}@tara.tsukuba.ac.jp
[2] National Institute of Informatics, Principles of Informatics Research Division
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan, onono@nii.ac.jp

ABSTRACT

This paper proposes a blind synchronization of ad-hoc microphone array in the short-time Fourier transform (STFT) domain with the optimized frame analysis centered at non-integer discrete time. We show that the drift caused by sampling frequency mismatch of asynchronous observation channels can be disregarded in a short interval. Utilizing this property, the sampling frequency mismatch and the recording start offset are estimated roughly by finding two pairs of the short intervals corresponding to the same continuous time. Using the estimate, STFT analysis is synchronized roughly between channels with optimized frame central. Since the optimized frame central is generally non-integer, we approximate the frame analysis by the linear phase filtering of the frame centered at the nearest integer sample. Maximum likelihood estimation refines the compensation of sampling frequency mismatch.

Index Terms— Ad-hoc microphone array, sampling frequency, maximum likelihood estimation, blind source separation

1. INTRODUCTION

Ad-hoc microphone array has attracted increasing attention in association with development of mobile recording devices [1]. By using simultaneous recording with separate mobile devices such as cell phones, voice recorders, video camera etc. as multichannel observation, microphone array technology obtains accelerated applicability. However, increased freedom brings various difficulty. Especially the asynchronous sampling of each channel causes drift of time, which seriously degrades time-difference analysis of array signal processing [2].

While many of the compensation methods of sampling frequency mismatch requires some prior information [1, 2, 3], we proposed fully blind compensation [4]. By disregarding small drift inside short-time frames, the compensation reduced to the small shift of the frames. To process very small time shift accurately, the compensation was conducted as linear phase filtering in the short-time Fourier transform (STFT) domain. Also the sampling frequency mismatch is estimated by maximum likelihood estimation. However, with long recordings the required shift becomes significantly large inside each frame, and the circular shift inside the frame fails to approximate the compensation. In this paper we propose an optimization method of frame analysis with non-integer frame shift. With correlation analysis, short interval pairs indicating the same continuous times in each channel is detected. By using two pairs of synchronous intervals, the synchronization is estimated roughly. The small error of the synchronized STFT analysis is compensated effectively by the maximum-likelihood linear phase.

2. PROBLEM STATEMENT

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. 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 $|\epsilon| \ll 1$. The relations between $x_i(n)$ and $\mathrm{x}_i(t)$ for $i = 1, 2$ are given by

$$x_1(n) = \mathrm{x}_1\left(\frac{n}{f_\mathrm{s}}\right), \quad n = 0, \ldots, N_1 - 1, \tag{1}$$

$$x_2(n) = \mathrm{x}_2\left(\frac{n}{(1+\epsilon) f_\mathrm{s}} + T_{21}\right), \quad n = 0, \ldots, N_2 - 1, \tag{2}$$

where N_i is the length of the digital signal $x_i(n)$. The relation between the synchronous pair of the discrete times n_1, n_2 of these two sampled signals $x_1(n_1)$, $x_2(n_2)$ correspond to the same continuous time t is given by a function $\phi_{21}(n)$ as

$$n_2 = \phi_{21}(n_1), \tag{3}$$

$$\phi_{21}(n) = (1+\epsilon)(n - D_{21}), \tag{4}$$

$$D_{21} = f_\mathrm{s} T_{21}, \tag{5}$$

where D_{21} stands for the discrete time of the first channel when the recording of the second channel starts. 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. With the integer values of n_1, the corresponding discrete times n_2 of the second channel are generally non-integer. Thus to obtain the signal $\hat{x}_2(n)$ of the second channel synchronized accurately to the first channel, the following infinite convolution of the sinc functions is required.

$$\begin{aligned}\hat{x}_2(n) &= x_2(\phi_{21}(n)) \\ &= \sum_{n'=-\infty}^{\infty} \mathrm{sinc}\left((1+\epsilon)(n - f_\mathrm{s} T_{21}) - n'\right) x_2(n'). \end{aligned} \tag{6}$$

Since such the infinite convolution cannot be operated in practice, some efficient approximation is necessary.

In this paper we discuss approximate synchronization of the asynchronous channels with blind estimation of D_{21} and ϵ with accuracy sufficient for array signal processing. Estimation accuracy

978-1-4799-0972-8/13/$31.00 ©2013IEEE

of ϵ is critical for array signal processing because the drift makes the time difference of arrival of each source time-varying. In contrast to the sampling frequency mismatch ϵ, estimation accuracy of D_{21} is not significant in specific classes of array signal processing such as blind source separation (BSS) [5] which do not use directions of arrival explicitly, and the error is accepted as long as it is much smaller than the frame length.

3. APPROXIMATE COMPENSATION OF SAMPLING FREQUENCY MISMATCH IN STFT DOMAIN

Since array signal processing is generally conducted in the STFT domain, we propose the STFT expression of the approximate synchronization between the channels.

3.1. Modeling sampling frequency mismatch in short-time frames

Before we proceed to STFT analysis, we discuss the effect of the 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 to the (n_1+m)-th sample of the first channel is given by the relation in (3) as

$$\begin{aligned}\phi_{21}(n_1+m) &= (1+\epsilon)(n_1 - D_{21}) + (1+\epsilon)m \\ &= \phi_{21}(n_1) + (1+\epsilon)m,\end{aligned} \tag{7}$$

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

$$\phi_{21}(n+m) \approx \phi_{21}(n) + m. \tag{8}$$

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 to be synchronous.

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

$$x_i^{\mathrm{fr}}(l, n_i) = w(l)\, x_i\left(l + n_i - \frac{L}{2}\right), \tag{9}$$

where $w(l)$ is an appropriate window function, is almost synchronous between the channels $i = 1, 2$. Since the sampling frequency mismatch ϵ is generally in the order of 10^{-5} and typical frame length of microphone array signal processing is in the order of 0.1 second, the largest approximation error $|\epsilon L/2|$ of the time, which appears in the beginning and the end of the frame with $m \approx \pm L/2$ in (7) and (8), is usually below the order of 1 µs in such a frame analysis. Note that the influence of the errors are reduced with typical choice of the window function $w(l)$ to suppress the amplitude near both ends.

3.2. Approximate synchronization in STFT domain

Here we discuss the STFT expression of the approximation of $x_2^{\mathrm{fr}}(l, n_2)$ by the linear phase filtering of the frame centered at the rounded integer sample. Although the linear phase gives the circular convolution inside the frame, its error can be disregarded with the large frame length $L \gg 1$.

The STFT analysis of the i-th channel of the frame centered at the sample n is given by

$$X_i(k, n) = \sum_{l=0}^{L-1} x_i^{\mathrm{fr}}(l, n) \exp\left(-\frac{2\pi \jmath k l}{L}\right), \tag{10}$$

where $k = -L/2, \ldots, L/2-1$ is the discrete frequency index. Note that the transform is calculated by fast Fourier transform in the practical processing. According to (3), the discrete time of the second channel synchronous to the central time n_1 of $X_1(k, n_1)$ is given by $n_2 = \phi_{21}(n_1)$. To approximate the STFT centered at the non-integer time $\phi_{21}(n_1)$, we apply the frame analysis with the nearest integer central time, and compensate the effect of the rounding by the circular time shift using the linear phase filter.

First we obtain the frame analysis $x_2^{\mathrm{fr}}(l, \lfloor \phi_{21}(n_1) \rceil)$ of the second channel $x_2(n)$ centered at the integer sample $\lfloor \phi_{21}(n_1) \rceil$ nearest to the desired central time $\phi_{21}(n_1)$, given by

$$\lfloor \phi_{21}(n_1) \rceil = \arg\min_n |\phi_{21}(n_1) - n|, \quad n \in \mathbb{Z}. \tag{11}$$

Since the central sample $\lfloor \phi_{21}(n_1) \rceil$ is delayed from the non-integer time $\phi_{21}(n_1)$ by $\phi_{21}(n_1) - \lfloor \phi_{21}(n_1) \rceil$, we obtain the approximation of synchronization in the STFT domain by compensating the delay with the linear phase filter as

$$\begin{aligned}&\hat{X}_2(k, \phi_{21}(n_1)) = \\ &\quad X_2(k, \lfloor \phi_{21}(n_1) \rceil) \exp\left(\frac{2\pi \jmath k (\phi_{21}(n_1) - \lfloor \phi_{21}(n_1) \rceil)}{L}\right).\end{aligned} \tag{12}$$

It is worth noting that the time domain signal $\hat{x}_2(n; \epsilon, D_{21})$ approximately synchronized to $x_1(n)$ can be given by the inverse STFT analysis of $\hat{X}_2(k, \phi_{21}(n_1))$ with the frame shift common to $X_1(k, n_1)$.

4. COARSE BLIND SYNCHRONIZATION

In [4] we proposed a blind estimation of sampling frequency mismatch ϵ and its compensation by the linear phase filtering of the STFT analyzed analogously to the other channel. However, with the long observation, the linear phase compensation is insufficient because the large gaps of the frame centrals. To solve this problem, we propose a roughly synchronized STFT analysis by estimating the recording start offset D_{21} and the sampling frequency mismatch ϵ roughly and substituting them in the STFT synchronization discussed in the previous section.

4.1. Estimation with two pairs of corresponding time

Before the discussion of estimation procedure, we analyze the condition that the parameters ϵ and D_{21} have to satisfy when two pairs $\{n_{\mathrm{A}1}, n_{\mathrm{A}2}\}$ and $\{n_{\mathrm{B}1}, n_{\mathrm{B}2}\}$ of synchronous times are given:

$$n_{\mathrm{A}2} = (1+\epsilon)(n_{\mathrm{A}1} - D_{21}), \tag{13}$$

$$n_{\mathrm{B}2} = (1+\epsilon)(n_{\mathrm{B}1} - D_{21}), \tag{14}$$

which give ϵ and D_{21} as

$$\epsilon = \frac{n_{\mathrm{B}2} - n_{\mathrm{A}2}}{n_{\mathrm{B}1} - n_{\mathrm{A}1}} - 1, \tag{15}$$

$$D_{21} = \frac{n_{\mathrm{A}1} n_{\mathrm{B}2} - n_{\mathrm{A}2} n_{\mathrm{B}1}}{n_{\mathrm{B}2} - n_{\mathrm{A}2}}. \tag{16}$$

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 ϵ and D_{21}. Although accurate estimation of the synchronized time pairs is difficult, even their rough estimation is useful for the coarse synchronization.

4.2. Estimating synchronous times with correlation analysis

We discuss the rough estimation of $n_{\mathrm{A}i}, n_{\mathrm{B}i}$ for $i = 1, 2$. The basic idea is that the nearly synchronous pairs of short intervals can be detected using correlation. Since the observation is the mixture of multiple sources with different time differences of arrival, the estimation of intervals has error and precise one cannot be obtained. Thus to reduce the effect of the error in the parameter estimation in (15) and (16), it is preferable to estimate $n_{\mathrm{A}i}$ of small values and $n_{\mathrm{B}i}$ of large values. To obtain the estimation of the intervals, we analyze the intervals with the amplitude exceeds the threshold near the beginning and the end of the recording. Since it is unknown which channel starts and ends earlier or later, the candidates of the short intervals are chosen from both channels near each of the beginning and the end, and choose the better one with the higher correlation to other channel.

First, as the candidates of $n_{\mathrm{A}i}, i = 1, 2$, we find the earliest samples $n'_{\mathrm{A}ij}, (i, j) = (1, 2), (2, 1)$ with the amplitude exceeding the threshold δ in the region over $N/2$ samples later than the beginning sample as

$$n'_{Aij} = \arg\min_{n} \{|x_i(n)| > \delta\} + N/2 - 1, \quad n \geq N/2, \tag{17}$$

where N shows the length of the short interval. Similarly, as the candidates of $n_{\mathrm{B}i}, i = 1, 2$, we find the latest samples $n'_{Bij}, (i, j) = (1, 2), (2, 1)$ with the amplitude exceeding the threshold δ in the region over $N/2 - 1$ samples earlier than the end of the channel's recording as

$$n'_{Bij} = \arg\max_{n} \{|x_i(n)| > \delta\} - N/2, \quad n < N_i - N/2. \tag{18}$$

Next, for each of $h = \mathrm{A}, \mathrm{B}$, $(i, j) = (1, 2), (2, 1)$, we find the sample n''_{hij} of $x_j(n)$ to maximize the correlation with the selected interval as

$$n''_{ij} = \arg\max_{n''} \sum_{n=-L/2}^{N/2-1} x_i\left(n + n'_{hji}\right) x_j\left(n + n''\right). \tag{19}$$

Although the direct calculation of this convolution is computationally complex, equivalent operation is efficiently computed by the FFT-based overlap-and-add convolution.

Subsequently, for each of $h = \mathrm{A}, \mathrm{B}$, we find the interval pair with higher the correlation between the channels. We select the one of the index pairs $(i, j) = (1, 2), (2, 1)$ with the higher coherence $|\gamma_h(i, j)|^2$ given by

$$(i^*_h, j^*_h) = \arg\max_{(i,j)=(1,2),(2,1)} |\gamma_h(i, j)|^2, \tag{20}$$

$$|\gamma_h(i, j)|^2 = \frac{\left| \sum_{n=-\frac{L}{2}}^{N-1} x_i\left(n + n'_{hji}\right) x_j\left(n + n''_{hij}\right) \right|^2}{\sum_{n=-\frac{L}{2}}^{N-1} \left|x_i\left(n + n'_{hji}\right)\right|^2 \sum_{n=-\frac{L}{2}}^{N-1} \left|x_j\left(n + n''_{hij}\right)\right|^2}. \tag{21}$$

Finally, $n_{\mathrm{A}1}, n_{\mathrm{A}2}, n_{\mathrm{B}1}$ and $n_{\mathrm{B}2}$ estimated by the interval pairs with the higher coherence as

$$n_{hi^*_h} = n'_{hj^*_h i^*_h}, \tag{22}$$

$$n_{hj^*_h} = n''_{hi^*_h j^*_h}. \tag{23}$$

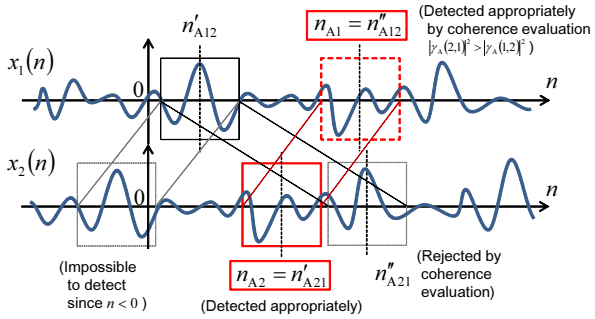

Figure 1: Estimation of $n_{\mathrm{A}1}$ and $n_{\mathrm{A}2}$ corresponding to the same continuous time near the beginning of the asynchronous recording.

The parameters ϵ and D_{21} are roughly estimated by substituting these in (15) and (16). Finally, these roughly-estimated parameters are used in the STFT synchronization discussed in the previous section to obtain.

5. FINE COMPENSATION OF SAMPLING FREQUENCY MISMATCH

Here we discuss the fine compensation of the sampling frequency mismatch ϵ. Since the estimation of D_{21} discussed in the previous section is reliable, we estimate the parameter ϵ' to modify the estimation of sampling frequency in the following form keeping the modification of the offset as small as possible.

$$\begin{aligned} n_2 &= \phi'_{21}\left(n_1; \epsilon'\right) \\ &= (1 + \epsilon)\left(\left(1 + \epsilon'\right)(n_1 - M) + M - D_{21}\right), \end{aligned} \tag{24}$$

where M denotes the central sample of the whole the duration of $x_1(n)$ the frame analysis is applied.

Since the rough estimation of ϵ is already obtained, the modification of synchronization by ϵ' tends to be small. Thus the fine synchronization is conducted as the linear phase filtering of $\hat{X}_2(k, \phi_{21}(n_1))$ in (12) to obtain the modified signal $\hat{X}'_2(k, \phi'_{21}(n_1; \epsilon', M))$ as

$$\begin{aligned} &\hat{X}'_2\left(k, \phi'_{21}\left(n_1; \epsilon'\right)\right) \\ &\quad = \hat{X}_2\left(k, \phi_{21}(n_1)\right) \exp\left(\frac{2\pi \jmath k \epsilon'(n_1 - M)}{L}\right). \end{aligned} \tag{25}$$

The parameter ϵ' can be optimized as a maximum likelihood estimate assuming the two-channel signal given by

$$\hat{\mathbf{X}}'\left(k, n_1; \epsilon'\right) = \left[X_1(k, n_1), \hat{X}'_2\left(k, \phi'_{21}\left(n_1; \epsilon'\right)\right)\right]^{\mathsf{T}} \tag{26}$$

is stationary with the correctly estimated mismatch modification ϵ' [4]. The log likelihood is given by

$$L\left(\epsilon'\right) = -\sum_{k} \log\det \sum_{n_1} \hat{\mathbf{X}}'\left(k, n_1; \epsilon'\right) \hat{\mathbf{X}}'\left(k, n_1; \epsilon'\right)^{\mathsf{H}}. \tag{27}$$

The parameter ϵ' can be optimized by the combination of discretized search and golden section search as in [4].

Note that the optimization of ϵ' is equivalent to the following update of ϵ and D_{21}.

$$\epsilon \leftarrow (1 + \epsilon)\left(1 + \epsilon'\right) - 1, \tag{28}$$

$$D_{21} \leftarrow \frac{\epsilon' M + D_{21}}{1 + \epsilon'}. \tag{29}$$

Table 1: Experimental conditions.

Signal length [s]	3, 5, 10, 20, 30, 60, 120, 300 and 600
Reverberation time	T_{60} of 130 ms
Frame length L	4,096 samples
Frame shift	2,048 samples
Source distances	1.5 m
Source directions	[-50°, 30°], [-60°, -10°]
Microphone spacing	2 cm
N	16,000 samples
Candidates of discretized search [4]	10 samples from $[-2 \times 10^{-4}, 2 \times 10^{-4}]$

Although the synchronization of $\hat{\mathbf{X}}'(k, n_1; \epsilon')$ in the STFT domain with the maximum likelihood estimate ϵ' is sufficiently accurate as we show in the experiment, it is also possible to apply other resampling methods to synchronize the channels with the estimated parameters ϵ and D_{21} in (28) and (29).

6. EVALUATION

To confirm the effectiveness of the proposed blind synchronization, 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.

The observed signals are made by convolution of measured impulse responses and speech signals, which are generated by concatenation of Japanese word utterances. 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 Hz, and to one channel we gave modifications of sampling frequency of ± 0.5, ± 1, ± 1.5 Hz, which are realistic as practical bias of sampling frequencies. To generate the artificial sampling frequency mismatch, we used resampling with the polyphase filters. We used auxiliary-function-based independent vector analysis [6] to conduct BSS. Other conditions are listed in Table 1.

To examine accuracy of the estimation of sampling frequency mismatch ϵ, we compared the root mean squared errors (RMSEs) of the conventional method [4], the coarse search discussed in Sect. 4 and the proposed method in Fig. 2. It can be clearly seen that the estimation accuracy improves with the increase of the observation length. The estimation accuracy is improved from the coarse estimate by the proposed method. The conventional and the proposed methods performs similarly.

We compared the source separation performance in Fig. 3. The evaluation criterion is signal-to-distortion ratio (SDR) [7]. The curve labeled as no mismatch shows the performance of the synchronized recording without the artificial mismatch, and its SDRs are highest. We can see the very low performance of the unprocessed signal and compensation of synchronization is necessary in this condition. The performance of the conventional method degrades the performance with the observation longer than 30 s even with the high accuracy similar to the proposed method despite the accurate estimation. Thus the optimization of frame analysis is necessary with such the long data. We evaluated the upper limit of the STFT synchronization by substituting true values of ϵ and D_{21} in (12), which shows small degradation of about 1 dB from the no mismatch. The performance of the proposed method has little degradation from the upper limit. Note that such the small error is kept with the observation longer than 600 ms. Therefore it is confirmed that the proposed synchronization successfully recovers the source separation performance.

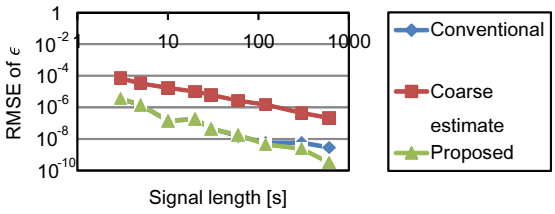

Figure 2: Root mean squared errors (RMSEs) of estimations of sampling frequency mismatches ϵ.

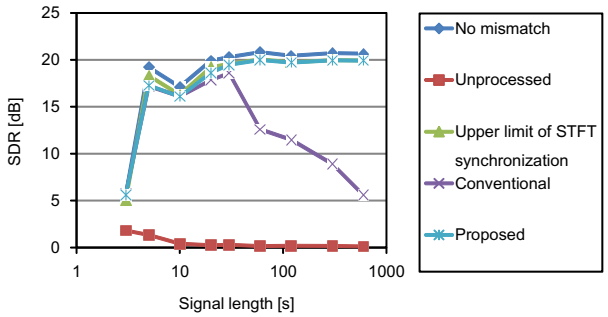

Figure 3: Source separation performance with SDR measure.

7. REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

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

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

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