Sampling Frequency Mismatch Estimation by Auxiliary-Function-Based Iterative Maximization of Double-Cross-Correlation

Kouei Yamaoka, Nobutaka Ono and Yukoh Wakabayashi

Graduate School of Systems Design, Tokyo Metropolitan University
6-6 Asahigaoka, Hino-shi, Tokyo 191-0065, Japan
{yamaoka-kouei@ed.,onono@,wakayuko@}tmu.ac.jp

***Abstract*—In this paper, we propose a new variant of sampling frequency mismatch (SFM) estimation based on double-cross-correlation processor (DXCP) by an auxiliary function method. SFM estimation is one of the key problems in signal processing on asynchronous microphone arrays. Previously, a DXCP was proposed for obtaining an accurate and robust SFM estimate. The DXCP estimates the SFM by maximizing a cross-correlation (CC) function, where parabolic interpolation is employed to attain the sub-sample time delay (STD) estimate between two observed signals. While, we previously proposed a highly accurate technique of STD estimation based on the auxiliary function method, which reaches a local maximum of a CC function, achieving a better result than the parabolic interpolation. In this paper, we thus extend the DXCP using our approach of STD estimation to improve the performance of SFM estimation. In experiments, we confirm that the proposed method shows a monotonic increase in objective function in the DXCP and achieves better performance than the original DXCP.**

***Index Terms*—sampling frequency mismatch, cross-correlation, auxiliary function, majorization–minimization**

I. Introduction

Microphone array signal processing realizes fundamental techniques such as sound source separation [1], [2] and source localization [3], [4]. These techniques, developed with an asynchronous distributed microphone array (ADMA) [5], are expected to improve the convenience and scope of application for audio technologies such as automatic speech recognition and acoustic scene classification systems. Many recording devices placed widely apart provide large-scale spatial information, which may also profit signal processing techniques exploiting spatial cues, i.e., time difference of arrival and variations in amplitudes.

One of the essential problems in ADMAs is sampling frequency mismatch (SFM) among individual recording devices. Traditional array signal processing is performed under the assumption that observed signals are completely synchronized; in other words, all microphones are installed on an identical A/D converter for sampling signals under exactly matched sensor clocks. However, in the case of using different devices, SFMs can exist among them, that will make the performance of the array signal processing degrade significantly. Therefore, the amount of SFM must be estimated in advance and compensated for post-stage processes.

Most techniques of SFM estimation are based on the maximization of a cross-correlation (CC) function between a reference signal and the signal to be compensated [6], [7]. A double-cross-correlation processor (DXCP) has recently been proposed for accurate and robust SFM estimation [8]. It shows notable performance compared with conventional SFM estimation [9]. In these techniques, there is a trade-off between estimation accuracy and computational complexity in the maximization of the CC. In general, the CC between two discrete signals is computed at every integer time delay, but its accuracy is insufficient for many applications, including SFM estimation. For attaining a sub-sample time delay (STD) estimate, several variations of interpolation have been proposed, such as parabolic interpolation [10], Gaussian curve fitting [11], and others [12]–[15]. In addition, it is possible to try to find the maximum of the continuous CC function directly. Following the Nyquist–Shannon sampling theorem [16], the continuous CC can be considered for band-limited signals. However, its maximization is a nonconvex problem, making it difficult to obtain the optimal solution. To estimate it, Miyabe et al. [6] and Wang and Doclo [7] used the golden-section search algorithm [17] and an exhaustive search scheme, respectively.

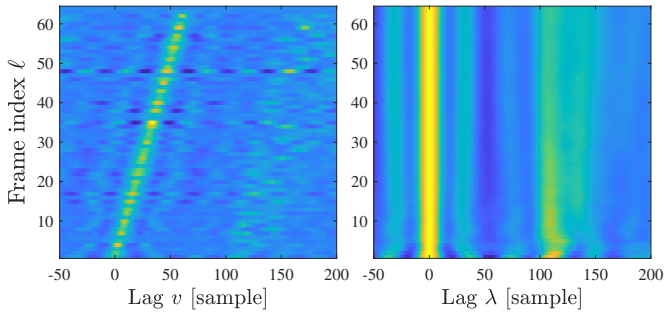
In contrast, we previously proposed a technique of maximizing a continuous CC via the auxiliary-function-based iterative updates for STD estimation [18]. Theoretically, this technique yields the same estimate as exhaustive search, but the computational cost is markedly low owing to efficient updates. Thus, in this study, we adapt this technique to the DXCP-based SFM estimation, which uses parabolic interpolation for CC maximization, to improve the estimation accuracy.

We conducted numerical experiments and confirmed the efficacy of the proposed method.

II. Sampling frequency mismatch estimation

A. Problem formulation

In this study, we consider estimating an SFM between two different microphones, i.e., the observed signals sampled by distinct A/D converters. Here, we assume two conditions: the sampling frequencies of these microphones almost match (the nominal value is the same or they are already resampled) and the sound source does not move. In a typical situation in ADMAs, the positions of microphones and sources are unknown. Therefore, we must estimate an SFM using only the observed signals.



(a) First CC. (b) Normalized second CC.

Fig. 1: Examples of the first and second CCs in DXCP for simulated signal with time-invariant SFM, where $M = 2 \times f_s$ and $B = 1 \times f_s$.

Let $\tilde{x}_1(t)$ and $\tilde{x}_2(t)$ be continuous signals observed by the microphones, and $x_1(n_1)$ and $x_2(n_2)$ be discrete signals sampled at unknown sampling frequencies f_{s_1} and f_{s_2}, respectively, where t denotes a continuous time variable and n_1 and n_2 denote sample indices. Without loss of generality, we consider $x_1(n_1)$ as a reference signal and represent its sampling frequency as f_s (i.e., $f_s = f_{s_1}$). Then, the relationships between the continuous and discrete signals are

$$x_1(n_1) = \tilde{x}_1\left(\frac{n_1}{f_s}\right), \tag{1}$$

$$x_2(n_2) = \tilde{x}_2\left(\frac{n_2}{(1+\epsilon)f_s} + \Delta T_{21}\right), \tag{2}$$

where ΔT_{21} denotes the delay of the recording start time of $\tilde{x}_2(t)$ with respect to $\tilde{x}_1(t)$ and can be more than several seconds. In this paper, we consider that ΔT_{21} is estimated and compensated sufficiently using a technique such as that proposed in [6]. ϵ represents the unknown SFM, defined as

$$\epsilon = \frac{f_{s_1}}{f_{s_2}} - 1, \tag{3}$$

where we assume that ϵ is time-invariant. Under this assumption, we can consider the linear phase drift model [6], which means that the phase difference between two signals changes linearly with time. Now, our goal is to estimate SFM ϵ from the observed signals.

B. Double-cross-correlation processor

In this section, we briefly introduce a DXCP [8]. The use of CC is typical in estimating an SFM in existing studies such as [6] and [7], while Chinaev et al. [8] addressed the consideration of the CC between the CCs of the two observations, namely, the double-cross-correlation (DXC). First, we perform windowed frame analysis as

$$x_{i,\ell}(m) = w(m)\ x_i((\ell-1)B + m - 1), \tag{4}$$

where $i \in 1, 2$ denotes the microphone index, $\ell = 1, \ldots, L$ denotes the frame index, and m denotes the sample index in each frame. $w(m)$ is an analysis window that takes a real value for $m = 1, \ldots, M$ and 0 for the other m, and B represents the frame shift length. Then, the CC $\phi(\ell, v)$ of the windowed signals is defined as

$$\phi(\ell, v) = \sum_{m=1}^{M} x_{1,\ell}(m)\ x_{2,\ell}(m+v), \tag{5}$$

where $v \in \{-\Gamma, \ldots, \Gamma\}$ and $0 < \Gamma \leq M - 1$ denotes a lag index. In this paper, we call $\phi(\ell, v)$ the first CC. An example of the first CC is shown in Fig. 1(a). We can see that the lag v that maximizes the first CC increases linearly with the frame index ℓ, which demonstrates the linear phase drift model [6].

We consider a second CC $\psi(\ell, \lambda)$ defined as

$$\psi(\ell, \lambda) = \sum_{v=-\Gamma}^{\Gamma} \phi(\ell, v)\ \phi(\ell-1, v+\lambda), \tag{6}$$

where $\lambda \in \{-\Lambda, \ldots, \Lambda\}$ and $0 < \Lambda \leq 2\Gamma$ is a lag index with $\ell \geq 2$. The second CC is the CC between the first CCs at the ℓth and $(\ell-1)$th frames. After process (6), the second CC is normalized with respect to the maximum value and is averaged across all past frames. An example of the second CC is shown in Fig. 1(b). In accordance with this figure, a clear peak at a certain lag can be seen for all frames, whereas the peak positions drifted with time in the first CC, as shown in Fig. 1(a). As a result, the DXCP achieves robust SFM estimation by using the second CC.

Finally, the SFM estimate $\hat{\epsilon}$ is obtained by maximizing the second CC $\psi(\ell, \lambda)$ in the last frame L as follows:

$$\hat{\lambda}_{\mathrm{d}}(L) = \arg\max_{\lambda}\ \psi(L, \lambda), \tag{7}$$

$$\hat{\epsilon} = \frac{\hat{\lambda}_{\mathrm{s}}(L)}{B}, \tag{8}$$

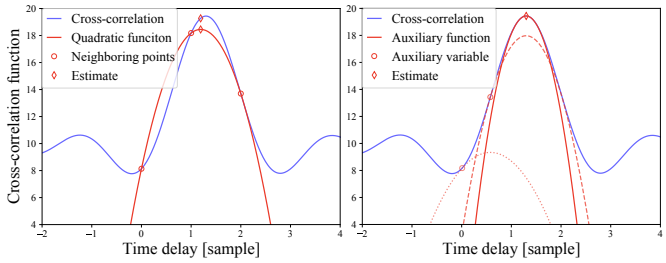
where λ_{d} and λ_{s} denote discrete and sub-sample lag estimates, respectively. In the original DXCP, it was proposed that λ_{s} is estimated via parabolic interpolation, i.e., λ_{s} corresponds to the vertex of the following quadratic function:

$$f(\lambda) = a\lambda^2 + b\lambda + c, \tag{9}$$

where the coefficients a, b, and c of the quadratic function $f(\lambda)$ are determined by the three neighboring points $\hat{\lambda}_{\mathrm{d}}(L)-1$, $\hat{\lambda}_{\mathrm{d}}(L)$, and $\hat{\lambda}_{\mathrm{d}}(L)+1$. By using the vertex of the quadratic function instead of the integer lag estimate λ_{d}, we can obtain better estimates, as shown in Fig. 2(a).

C. Discussion for better estimation with DXCP

Chinaev et al. [8] reported that the accuracy of the SFM estimate $\hat{\epsilon}$ via the DXCP strongly depends on the window length M. The optimum M was 10 s in their experiments [8]. A larger or smaller frame length degrades performance. The reason for this degradation can be as follows: in the case of a smaller frame length, the lag between two windowed frames becomes larger relative to the frame length; thus, the two frames do not match. In the case of a larger frame length, the time variation of the lag within a frame becomes non-negligible. Both cases decrease the correlation between two frames. Because discussion about the degradation due to this reason is not within the scope of this paper, please refer to [6], [7] for details.



(a) Parabolic interpolation (b) Proposed method

Fig. 2: Examples of CC maximization via (a) parabolic interpolation with three neighboring points and (b) proposed auxiliary-function-based iterative updates. The CC function was calculated for every 0.001 samples for smooth plotting.

Another reason for the performance dependence on the frame length is that the frame length affects the accuracy of peak estimation in DXCP. In this paper, we focus on how to attain the peak. Since the CC between two discrete signals is typically computed at every discrete lag index, the accuracy depends on the sampling frequency and the frame length M. In the original DXCP [8], the use of parabolic interpolation [10] in the vicinity of the maximum of the second CC to estimate the STD, which is an approximate solution (see Fig. 2(a)), has been proposed. The computational cost of this approach is low; however, the accuracy still depends on the sampling frequency. In contrast, we previously proposed the auxiliary-function-based approach to maximize a CC [18]. We thus adopt this approach to the DXCP to improve the estimation accuracy.

Other means to improve the SFM accuracy include the sharpening of the peak of CC using generalized cross-correlation phase transform (GCC-PHAT) [19], [20] and performing SFM estimation and signal compensation iteratively [7], [8], [21]. The combination of such extensions and the proposed method in this paper will be considered in the future.

III. Maximization of Cross-Correlation Function via Auxiliary-Function-Based Iterative Updating

In this section, we explain the technique of maximizing the second CC by the auxiliary function method. In subsections III-A–III-D, we briefly introduce the auxiliary-function-based STD estimation for completeness, proposed in [18].

A. Problem formulation

First, for discrete signals $x_1(n_1)$ and $x_2(n_2)$, we assume that the effect of an SFM can be ignored in a short-time frame and that $n = n_1 = n_2$. The second CC $\psi(L, \lambda)$ (hereafter, we omit the index L for simplicity) is represented as the sum of complex sinusoids using the discrete Fourier transform (DFT) coefficients. We here assume that signals $x_1(n_1)$ and $x_2(n_2)$ are strictly band-limited and follow the Nyquist–Shannon sampling theorem [16]. Then, the discrete lag λ can be replaced by the continuous lag $t \in \mathbb{R}$, and the second CC can be rewritten as

$$\psi(t) = \frac{1}{N} \sum_{k=-\Lambda}^{\Lambda} S_k e^{j2\pi kt/N}, \tag{10}$$

where $N = 2\Lambda + 1$ is the length of $\psi(\lambda)$ ($N = 4M - 3$ at maximum), k denotes a discrete frequency index, and S_k is the DFT of $\psi(\lambda)$, i.e., the cross-spectrum of the first CCs $\phi(\ell)$ and $\phi(\ell - 1)$. Note that $\psi(t) = \psi(\lambda_{\mathrm{d}})$ when t is an integer. We consider finding the continuous time variable $t \in \mathbb{R}$ maximizing the continuous function (10). It can be rewritten as a sum of cosines using the conjugate symmetry of S_k,

$$\psi(t) = \sum_{k=0}^{\Lambda} A_k \cos(\omega_k t + \phi_k), \tag{11}$$

where $A_k = \frac{\beta_k}{N}|S_k|$, $\omega_k = 2\pi\frac{k}{N}$, $\phi_k = \angle S_k$ and $\beta_0 = 1$ and $\beta_k = 2$ $(k \neq 0)$. Our goal is to compute the sub-sample lag estimate

$$\hat{t} = \arg\max_{t \in \mathbb{R}} \psi(t). \tag{12}$$

B. Proposed sub-sample time delay estimation

The auxiliary function method (also known as the majorization-minimization (MM) algorithm [22]) is a well-known method, and various application have been proposed [23]–[25]. In our problem, an auxiliary function $Q(t, \boldsymbol{\theta})$ that satisfies the following is required:

- $\psi(t) \geq Q(t, \boldsymbol{\theta})$ for any t and $\boldsymbol{\theta}$,
- for any t, $\exists \boldsymbol{\theta} = g(t)$ such that $\psi(t) = Q(t, \boldsymbol{\theta})$,

where $\boldsymbol{\theta} = (\theta_0, \theta_1, \cdots, \theta_{N/2})$ are auxiliary variables. Provided such a $Q(t, \boldsymbol{\theta})$ exists, and given an initial estimate $\hat{t}^{(0)}$, the following sequence of updates with the iteration index q is guaranteed to converge to a local maximum:

$$\boldsymbol{\theta}^{(q)} = g(\hat{t}^{(q)}), \quad \hat{t}^{(q+1)} = \arg\max_{t \in \mathbb{R}} Q(t^{(q)}, \boldsymbol{\theta}^{(q)}). \tag{13}$$

C. Quadratic auxiliary function for continuous cross-correlation function

This section provides an auxiliary function for $\psi(t)$ [18].

Theorem 1. *The following is an auxiliary function for $\psi(t)$,*

$$Q(t, \boldsymbol{\theta}) = \sum_{k=0}^{N/2} \left\{ -\frac{A_k}{2} \frac{\sin\theta_k}{\theta_k} (\omega_k t + \phi_k + 2n_k\pi)^2 + \mathrm{C}_k \right\}, \tag{14}$$

where C_k is a constant term that does not include t and $n_k \in \mathbb{Z}$ is such that $|\omega_k t + \phi_k + 2n_k\pi| \leq \pi$. The auxiliary variables are θ_k and n_k, then $Q(t, \boldsymbol{\theta}) = \psi(t)$ when

$$\theta_k = \omega_k t + \phi_k + 2n_k\pi. \tag{15}$$

This theorem is a direct consequence of the following inequality for a cosine function [18].

Proposition 1. *Let $|\theta_0| \leq \pi$. For any real number θ, the following inequality is satisfied:*

$$\cos\theta \geq -\frac{1}{2}\frac{\sin\theta_0}{\theta_0}\theta^2 + \left(\cos\theta_0 + \frac{1}{2}\theta_0 \sin\theta_0\right). \tag{16}$$

When $|\theta_0| < \pi$, equality holds if and only if $|\theta| = |\theta_0|$. When $|\theta_0| = \pi$, equality holds if and only if $\theta = (2n+1)\pi$, $n \in \mathbb{Z}$.

Proof of theorem 1 and proposition 1 are detailed in [18].

D. Derivation of auxiliary function and update rules

Since $Q(t, \boldsymbol{\theta})$ is a quadratic function, it is easily maximized with respect to t by putting its derivative to zero:

$$\frac{\partial Q}{\partial t} = -\sum_{k=0}^{N/2} A_k \omega_k \frac{\sin\theta_k}{\theta_k}(\omega_k t + \phi_k + 2n_k\pi) = 0. \quad (17)$$

Then, under the condition of equality (15), we can substitute $\phi_k + 2n_k\pi = \theta_k - \omega_k t$ and obtain the following update rules:

$$n_k^{(q)} \leftarrow \arg\min_{n\in\mathbb{Z}} \left|\omega_k t^{(q)} + \phi_k + 2n\pi\right|, \quad (18)$$

$$\theta_k^{(q)} \leftarrow \omega_k t^{(q)} + \phi_k + 2n_k^{(q)}\pi, \quad k = 0, \ldots, \frac{N}{2}, \quad (19)$$

$$t^{(q+1)} \leftarrow t^{(q)} - \frac{\sum_{k=0}^{N/2} A_k \omega_k^2 \left(\sin\theta_k^{(q)}/\theta_k^{(q)}\right)\frac{\theta_k^{(q)}}{\omega_k}}{\sum_{k=0}^{N/2} A_k \omega_k^2 \left(\sin\theta_k^{(q)}/\theta_k^{(q)}\right)}. \quad (20)$$

E. Adaptation for DXCP

To use our approach in the postprocessing of the DXCP, we have two choices in selecting the initial estimate $t^{(0)}$. The first one uses the discrete lag index that maximizes the second CC, i.e., the result of (7). The second one is the parabolic interpolation, which is the original method proposed in [8]. Basically, the better initial estimate leads to faster convergence, while the convergence point is the same owing to the property of the auxiliary function method. With the initial estimate $t^{(0)}$, we can obtain the final sub-sample precision lag estimate $\hat{t} = \hat{\lambda}_{\mathrm{s}}$ by iteratively updating the lag estimate and auxiliary variables using (18)–(20), where $\hat{\lambda}_{\mathrm{s}}$ maximizes the objective function (11). Finally, we compute (8) with $\hat{t}$ to obtain the SFM estimate.

IV. Experimental evaluation

To evaluate the efficacy of the proposed method, we investigated the performance of SFM estimation in terms of the convergence, as described in subsection IV-A, and the estimation accuracy, as described in subsection IV-B. In these experiments, we used microphone 1 as the reference microphone with a sampling rate of 16 kHz. Then, we estimated the unknown sampling frequency at microphone 2. For analysis window length M, we tested 12 variations, namely, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 15, and 20 s. Window shift length B was set to half of M. We used a Hann window for the windowed framing.

A. Convergence

In this experiment, we used Japanese male speech sampled at 16 000 Hz as a source signal. The signal of microphone 2 was resampled at 16 010 Hz, which is an unrealistic scenario, and was simulated merely for verifying the convergence in this experiment. The initial estimate $t^{(0)}$ used in the proposed method was a discrete sample that maximizes the second CC. We show the convergence of the objective function over frame length and initial estimates in Figs. 3 and 4, respectively.

Fig. 3 indicates that the proposed method shows the monotonic increase in the objective function for any frame length. Empirically, when the frame length is large, the objective function exhibits a gentle curvature. Therefore, the proposed

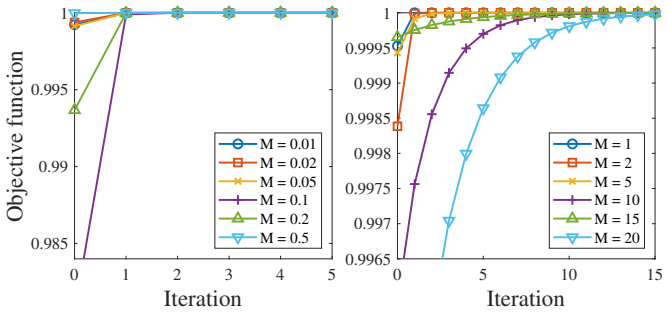


Fig. 3: Objective function (second CC) and state of its convergence over the frame length M.

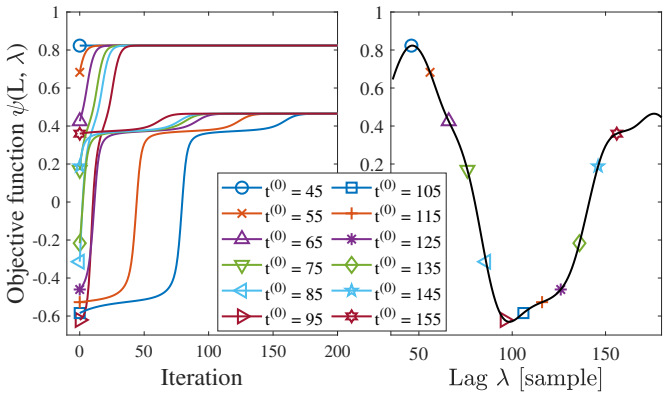


Fig. 4: Objective function (second CC) and state of its convergence for the initial estimates $t^{(0)}$ of every 10 samples. Each circle represents the position of the initial estimate , where the leftmost circle corresponds to the discrete lag maximizing the objective function.

method requires more iterations for convergence, but a better SFM estimate is obtained, as discussed in the next section. Conversely, the small frame length causes an acute peak of the second CC, which leads to faster convergence. Note that the shape and range of the objective function vary depending on the frame length. Thus, to align the range and display all curves in these two figures, we normalized the value of the objective function at each iteration so that the value at the last 20th iteration becomes one.

Fig. 4 shows that the proposed method with arbitrary initial estimates converged to the local maximum, when the frame length was 10 s. We can confirm that the proposed method reaches the global maximum by selecting the initial estimate from the unimodal period containing the global maximum. Even if the initial estimate is outside that unimodal period, convergence to the local maximum is guaranteed. This is one of the essential properties of the auxiliary function method. The better the initial estimate, the faster the convergence is.

B. Estimation accuracy

We used eight types of sound source, namely, two samples of male/female Japanese/English speeches, with a signal length of 25 s. The signal at microphone 2 was resampled by the accurate sinc-interpolation with a given SFM ϵ. ϵ was randomly selected from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ with variance $\sigma^2 = 31.25^2$ ppm and mean μ, where $\mu = 62.5$ ppm was used for half of the data and $\mu = -62.5$ ppm for the

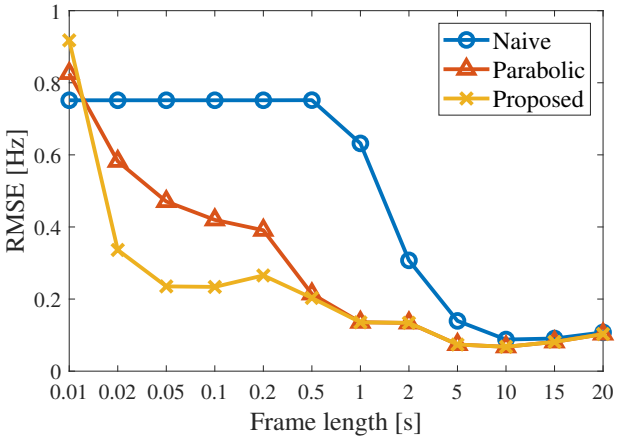


Fig. 5: Estimation error for sampling frequency mismatch as a function of the frame length.

remaining. Note that ϵ between signals sampled at 16 000 Hz and 16 001 Hz is 62.5 ppm. In this experiment, we used the DXCP with three types of postprocessing. "Naive" represents the DXCP with no postprocessing, that is, the estimate is a discrete sample maximizing the discrete second CC. "Parabolic" represents the original DXCP using parabolic interpolation. Finally, "Proposed" corresponds to the proposed method, i.e., the DXCP with auxiliary-function-based CC maximization. The number of iteration was 20.

Fig. 5 shows the estimation error of SFM. The horizontal axis shows the frame length, and the vertical axis shows the root mean square error (RMSE) between the sampling frequencies at microphone 2 (f_{s_2}) and its estimate ($\hat{f}_{s_2}$), where f_{s_2} and $\hat{f}_{s_2}$ were computed using the given ϵ and the ϵ estimated by using (3), respectively. The large frame length shows better RMSE than the small frame length because of the computation of ϵ in (8). The large B reduces the effect of error in the lag estimation (the maximization of the second CC). Thus, all methods attain almost the same performance as the frame length increases. The best accuracy is obtained when the window length is 10 s, where the RMSE is 0.067 52 Hz, and the error of SFM is 4.22 ppm. When the frame length is relatively small, the performance of "Naive" and the original DXCP decreases. In contrast, the proposed method still achieves the highest performance. This means that the proposed method yields a better solution for maximizing the second CC. In accordance with Fig. 5, the original DXCP requires a frame length of at least 0.5 s to obtain a better estimate. On the other hand, the proposed method requires a frame length of about 0.05 s to attain almost the same performance, which improves the practicality of the DXCP. From the above results, we confirmed the efficacy of the proposed method for SFM estimation.

V. Conclusions

In this paper, we present an efficient method for improving the DXCP-based SFM estimation. We demonstrate that the second CC can be lower-bounded at any point by the proposed auxiliary function. We then apply the auxiliary-function-based STD estimation, instead of the parabolic interpolation, for maximizing the second CC as postprocessing of DXCP. In the experiments, we confirmed that the proposed method converges to a local maximum with arbitrary initial estimates and achieves a better SFM estimate. In particular, the proposed method shows significant improvement for a small frame length, improving the practicality of the DXCP.

References

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

[2] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix fuctorization," *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[3] A. Beck, P. Stoica, and J. Li, "Exact and approximate solutions of source localization problems," *Signal Process.*, vol. 56, no. 5, pp. 1770–1778, Apr. 2008.

[4] T.-K. Le and N. Ono, "Closed-form and near closed-form solutions for TDOA-based joint source and sensor localization," *Signal Process.*, vol. 65, no. 5, pp. 1207–1221, Dec. 2016.

[5] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *Proc. SCVT*, pp. 1–6, 2011.

[6] S. Miyabe, N. Ono, and S. Makino, "Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation," *Signal Process.*, vol. 107, pp. 185–196, Feb. 2015.

[7] L. Wang and S. Doclo, "Correlation maximization-based sampling rate offset estimation for distributed microphone arrays," *IEEE/ACM Trans. ASLP*, vol. 24, no. 3, pp. 571–582, 2016.

[8] A. Chinaev, P. Thüne, and G. Enzner, "A double-cross-correlation processor for blind sampling rate offset estimation in acoustic sensor networks," in *Proc. ICASSP*, pp. 641–645, 2019.

[9] S. Markovich-Golan, S. Gannot, and I. Cohen, "Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming," in *Proc. IWAENC*, pp. 1–4, 2012.

[10] G. Jacovitti and G. Scarano, "Discrete time techniques for time delay estimation," *IEEE Trans. Signal Process.*, vol. 41, no. 2, pp. 525–533, Feb. 1993.

[11] L. Zhang and X. Wu, "On the application of cross correlation function to subsample discrete time delay estimation," *Dig. Signal Process.*, vol. 16, no. 6, pp. 682–694, Nov. 2006.

[12] F. Viola and W. F. Walker, "Computationally efficient spline-based time delay estimation," *IEEE Trans. UFFC*, vol. 55, no. 9, pp. 2084–2091, Sept. 2008.

[13] B. Qin, H. Zhang, Q. Fu, and Y. Yan, "Subsample time delay estimation via improved GCC PHAT algorithm," in *Proc. ICSP*, pp. 2579–2582, Oct. 2008.

[14] R. Tao, X.-M. Li, Y.-L. Li, and Y. Wang, "Time-delay estimation of chirp signals in the fractional Fourier domain," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2852–2855, July 2009.

[15] V. Martin, K. Jelena, and G. V. K, *Foundations of Signal Process.* Cambridge Univ. Press, Oct. 2014.

[16] C. E. Shannon, "Communication in the presence of noise," *Proc. IRE*, vol. 37, pp. 10–21, Feb. 1949.

[17] J. C. Kiefer, "Sequential minimax search for a maximum," *Proc. American Mathematical Society*, vol. 4, no. 3, pp. 502–506, 1953.

[18] K. Yamaoka, R. Scheibler, N. Ono, and Y. Wakabayashi, "Sub-sample time delay estimation via auxiliary-function-based iterative updates," in *Proc. WASPAA*, pp. 125–129, Oct. 2019.

[19] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[20] I. J. Tashev, *Sound Capture and Processing*, ser. Practical Approaches. John Wiley & Sons, Jul. 2009.

[21] J. Schmalenstroeer, J. Heymann, L. Drude, C. Boeddecker, and R. Haeb-Umbach, "Multi-stage coherence drift based sampling rate synchronization for acoustic beamforming," in *Proc. MMSP*, pp. 1–6, 2017.

[22] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The Am. Stat.*, vol. 58, no. 1, pp. 30–37, 2004.

[23] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NeurIPS*, pp. 556–562, 2000.

[24] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *Proc. ICASSP*, pp. 3437–3440, 2009.

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