Estimation of Consistent Time Delays in Subsample via Auxiliary-Function-Based Iterative Updates

Kouei Yamaoka, Student Member, IEEE, Yukoh Wakabayashi, Member, IEEE, and Nobutaka Ono, Senior Member, IEEE

Abstract—In this paper, we propose a new algorithm for the estimation of multiple time delays (TDs). Since a TD is a fundamental spatial cue for sensor array signal processing techniques, many methods for estimating it have been studied. Most of them, including generalized cross correlation (GCC)-based methods, focus on how to estimate a TD between two sensors. These methods can then be easily adapted for multiple TDs by applying them to every pair of a reference sensor and another one. However, these pairwise methods can use only the partial information obtained by the selected sensors, resulting in inconsistent TD estimates and limited estimation accuracy. In contrast, we propose joint optimization of entire TD parameters, where spatial information obtained from all sensors is taken into account. We also introduce a consistent constraint regarding TD parameters to the observation model. We then consider a multidimensional CC (MCC) as the objective function, which is derived on the basis of maximum likelihood estimation. To maximize the MCC, which is a nonconvex function, we derive the auxiliary function for the MCC and design efficient update rules. We additionally estimate the amplitudes of the transfer functions for supporting the TD estimation, where we maximize the Rayleigh quotient under the non-negative constraint. We experimentally analyze essential features of the proposed method and evaluate its effectiveness in TD estimation. Code will be available at https://github.com/onolab-tmu/AuxTDE.

I. INTRODUCTION

A time delay (TD) or time difference of arrival (TDOA) [1] observed between two sensors is a fundamental spatial cue for many signal processing techniques such as source localization, speech enhancement, and signal separation. Source localization and direction of arrival (DOA) estimation are essential techniques in audio [2]–[6] and other various engineering fields including, sonar [7], radar [8], ground-penetrating radar [9], and reflection seismology [10]. TD-based localization methods are widely studied because of its importance [11]–[13]. Speech enhancement is also important to extract an desired signal from noisy observation(s) [14]–[16]. Resampling [17], [18] and synchronization [19] are also general topics, especially for asynchronous distributed systems [20]. For these techniques, the TDs are important spatial features, where any improvement in TD estimation (TDE) directly translates to their better performance, and numerous studies exist [1], [21].

The mainstream TDE techniques are based on the generalized cross correlation (GCC) method [22], which is the most popular technique. Many techniques to improve the accuracy of GCC-based TDE have been proposed [23]–[25], where many studies focus on how to estimate a single TD between two different sensors. This is because we can easily adapt these techniques for an array of more than three sensors, where there are multiple TDs to be estimated, by applying them repeatedly for every pair of a reference sensor and another one. We call this simple solution the pairwise method in this paper.

Although the pairwise method is broadly used, we consider some problems. First, the cross correlation (CC) between a sensor and another far from it may be low, which degrades the accuracy of a GCC-based method. This may be a serious problem, especially in a large-scale environment, e.g., distributed systems [20]. Second, TDs estimated by the pairwise method are inconsistent. Assuming three sensors as shown in Fig. 1, the TD between sensors 1 and 3 is theoretically equal to the sum of those between sensors 1 and 2 and 2 and 3. However, this does not hold true because of, for example, estimation errors caused by the presence of noise. The best sensor to use as the reference sensor and how to identify it are unclear. Summarizing the above, the pairwise method only uses partial information obtained from a pair of sensors. For better TDE, the spatial information obtained from all sensors should be taken into account. In this paper, we thus aim to develop a new method for simultaneously estimating consistent TDs to improve their estimation accuracy.

Let us consider observing a source signal with an M channel sensor array. The number of sensor pairs is $M C_2$ (M choose 2), and the same number of TDs can be computed while only $M - 1$ TDs exist theoretically; in other words, there is redundancy. Some techniques utilize this redundancy...
to estimate a more accurate TD, e.g., by introducing a multichannel CC coefficient [1], [26], [27]. In contrast, we introduce a consistent constraint for estimating $M - 1$ TDs and consider multidimensional CC (MCC) that encodes spatial cues obtained from all sensors. Here, TDE via the maximization of the MCC has two difficulties: how to attain the TD estimates with subsample precision and how to estimate them efficiently.

The methodology of subsample TDE is widely studied to improve the accuracy of GCC-based estimation [22]–[24]. A naïve TD estimate is given by the location of the maximum of the discrete CC function between two sensors. Without further processing, the accuracy of a GCC-based method is limited by the sampling frequency. This can be a serious problem, especially for compact arrays. For example, for a sound source radiated in an ambient atmosphere, the maximum TD observed by two microphones spaced by 4 cm is less than 0.12 ms, i.e., less than two samples at 16 kHz. Such a sample level accuracy is insufficient for many applications, including, but not limited to, audio and optical processing techniques. In the case of two sensors, interpolation is a popular and effective method to attain the subsample TD estimate, where the CC function is interpolated in the vicinity of the maximum. The parabolic interpolation [28] determines a quadratic function whose curve goes through three neighboring points, namely, the discrete maximum point of the GCC function and its two adjacent points on both sides. By using the vertex of the quadratic function instead of the discrete maximum, we can obtain a subsample TD estimate. Various schemes have been proposed, e.g., Gaussian curve fitting [29] and others [30]–[32]. Yet another interpolation method is zero padding in the frequency domain, which corresponds to Dirichlet kernel interpolation [33], where the ratio of nonpadded to padded signal lengths is the attainable subsample accuracy. However, these techniques attain not the maximum of the GCC function but an approximate one and are only applicable to the pairwise method.

It is possible to find the maximum of the continuous GCC function directly. For band-limited signals, on the basis of the Nyquist–Shannon sampling theorem [34], [35], the continuous GCC function is obtained by sinc-interpolation of its discrete counterpart. Its maximization is a nonconvex problem without a known closed-form solution. Nevertheless, a locally optimal solution can be found with a search algorithm such as the exhaustive search scheme [36] and golden-section search (GSS) [37]. The GSS is an efficient algorithm of ternary searches, and thus, it must be performed on a two-dimensional parameter space selected from $(M - 1)$-dimensional parameter space; however, it is computationally demanding.

In contrast, we previously proposed a technique of maximizing a continuous CC function via the auxiliary-function-based iterative updates for subsample TDE [38]. This technique theoretically yields the same estimate as the exhaustive search, but the computational cost is markedly low owing to efficient updates. By extending this method, in this paper, we propose an efficient algorithm, namely, auxiliary-function-based TDE (AuxTDE), for obtaining highly accurate and consistent TD estimates. First, we derive the objective function, i.e., the MCC in the maximum likelihood (ML) sense. Then, we show that the objective function can be globally bounded by a quadratic auxiliary function and can then be repeatedly maximized for guaranteed convergence to a local maximum. Finally, we propose the AuxTDE algorithm to estimate consistent TDs. This method reaches the exact peak of the objective function, which means that the highly accurate subsample estimates can be obtained, and is the reference-free algorithm, which means that consistent TDs are obtained.

The rest of this paper is organized as follows. In section II, we define the signal model and briefly introduce the GCC-method. We also define what is consistent TDs. In section III, we formulate the estimation of consistent TDs and define the objective function. In section IV, we first explain the fundamental idea and theories for our problem. We here consider the case of two sensors as the simplest scenario, where we show the relationship between the objective function and CC. In section V, we generalize the algorithm described in section IV and propose the technique for estimating consistent TDs. Experimental analysis and numerical experiments are the topics in sections VI and VII, respectively. Section VIII concludes this paper.

Note that this paper is partially based on conference paper [38] in which we proposed the technique introduced in section IV. The contribution of this paper is that we extend the proposed method in [38] to multiple TDs and introduce the concept of consistent TDs.

II. TIME DELAYS ESTIMATION

A. Signal Model

In this paper, we consider estimating every interchannel subsample TD observed by an $M$ channel sensor array. Let $x_{mnk}$ be the short-time Fourier transform (STFT) representation of the signal observed by the $m$th sensor at the discrete frequency $k$ in the $n$th time frame. We model the observations as

$$x_{kn} = \tilde{s}_{kn}h_k + u_{kn},$$

$$h_k = [\hat{a}_{1k}e^{-i\omega_k t_1} \ldots \hat{a}_{Mk}e^{-i\omega_k t_M}],$$

where $\tilde{s}_{kn}$ is a source signal and $u_{kn} = [u_{1kn} \ldots u_{Mkn}]^T$ is noise signals at each sensor, where the superscript $T$ denotes nonconjugate transposition. $i = \sqrt{-1}$ denotes the imaginary unit, $\omega_k = 2\pi k/T$ denotes the normalized angular frequency, and $T$ denotes the number of samples in a frame. $h_k$ is the transfer function from the signal source to the sensor array, where $\hat{a}_{mk}$ and $t_m$ denote the amplitude and the time of arrival (TOA), respectively. The TOA represents the absolute time when the wave propagates to the sensor from the signal source.
We here consider the relative transfer function (RTF) \( g_k \) [39], [40], which is defined as the ratio of the transfer function \( h_k \):

\[
g_k(a_k, \tau) = \frac{h_k}{h_{rk}},
\]

\[
= [a_{1k}e^{-j\omega_1\tau_1} \cdots a_{MK}e^{-j\omega_M\tau_M}]^T,
\]

\[
a_k = [a_{1k} \cdots a_{MK}]^T,
\]

\[
\tau = [\tau_1 \cdots \tau_M]^T,
\]

where \( a_{mk} = \tilde{a}_{mk}/\tilde{a}_{rk} \) is the frequency-dependent relative amplitude (\( a_{mk} \in \mathbb{R}^+ \)), and \( \tau_{rm} = t_m - t_r \) is the continuous TD (TDOA) between the reference sensor \( r \) and the \( m \)th sensor. Therefore, \( a_{rk} = 1 \) for all \( k \), and \( \tau_{rr} = 0 \). This means that degree of freedom of the TDs is \( M - 1 \) (\( M \) TOAs minus one time origin). Without loss of generality, we set \( r = 1 \) in this paper. Finally, the signal model (1) with the RTF is

\[
x_{kn} = s_{kn}g_k(a_k, \tau) + u_{kn},
\]

where \( s_{kn} = \tilde{s}_{kn}/h_{rk} \) is the source image observed at the reference sensor. Then, the objective of this paper is to estimate TDs \( \tau \) from the observations \( x_{kn} \). In the rest of this paper, we denote scalars by regular letters and denote vectors and matrices by bold lower and upper case letters, respectively.

### B. GCC-based Time Delay Estimation

The GCC method is easily applicable to estimating multiple TDs; namely, we repeatedly obtain the GCC function by computing (7) and maximize it by solving (8) for all \( m \) except for \( r \). We call this approach, which uses only partial information, the pairwise method in this paper. The algorithm of the pairwise method is quite simple, and many existing methods can be adapted; however, these TD estimates computed by the pairwise method are basically inconsistent.

#### C. Consistent Time Delays

With \( M \) sensors and one signal source, there are \( M - 1 \) TDs. This means that although we need to select one reference sensor to determine the absolute time origin, the theoretical TDs are independent of its selection. In other words, the TDs should be consistent. Now, we define what are consistent TDs.

**Definition 1** (Consistent time delays). Let \( \tau_{rm} \) be the time delay between \( m \)th \((m = 1, \ldots, M) \) and \( r \)th sensors. The time delays are said to be consistent if

\[
\tau_{r'm} = \tau_r' + \tau_{rm} \quad \forall m, r, r',
\]

where \( r' \) is another reference sensor and \( r' = 1, \ldots, M \).

Definition 1 is always satisfied when \( \tau_{rm} = t_m - t_r \) in theory. However, there is no guarantee that TDs estimated by the pairwise method satisfy (10) since they estimate each TD separately; hence, the TD estimates are inconsistent. In this paper, we model the observed signal (6) that based on the consistent TDs, i.e., \( \tau_{rm} = t_m - t_r \). Thus, by estimating all TDs simultaneously, those estimates naturally satisfy (10).

### III. Problem Formulation

Starting from the signal model (6), we aim to estimate consistent TDs in the ML sense. We here assume that the noise signals \( u_{kn} \) follow the complex multivariate Gaussian distribution \( \mathcal{N}_c(\mu, I) \) with the mean \( \mu = 0 \) and variance of an \( M \times M \) identity matrix \( I \). First, we eliminate the variable \( s_{kn} \) by replacing it with the ML estimate \( \hat{s}_{kn} \). Given \( g_k(a_k, \tau) \), we obtain an ML estimate of \( s_{kn} \) as follows:

\[
\hat{s}_{kn} = \text{arg min}_s ||x_{kn} - sg_k(a_k, \tau)||_2^2
\]

\[
= \frac{g_k^H(a_k, \tau)x_{kn}}{g_k^H(a_k, \tau)g_k(a_k, \tau)},
\]

where \( \| \cdot \|_2 \) is the Euclidean norm of a vector and the superscript \( \mathbf{H} \) denotes conjugate transposition. We can also consider that \( u_{kn} \) follows the distribution \( \mathcal{N}_c(\mu, \Sigma) \) with the channel-dependent variance \( \Sigma \); \( \Sigma \) is an \( M \times M \) diagonal matrix whose \( m \)th diagonal entry is \( \sigma_m^2 \). In this case, by using the following weighted vectors instead of the original ones, we obtain the same ML estimate \( \hat{s}_{kn} \):

\[
x_{kn}' = \text{diag}(\sigma)^{-1}x_{kn},
\]

\[
a_k' = \text{diag}(\sigma)^{-1}a_k,
\]

where \( \text{diag}(\cdot) \) is a function that returns a square diagonal matrix with the elements of an input vector and \( \sigma = [\sigma_1 \cdots \sigma_M] \). Hereafter, we basically consider the channel-dependent case and omit prime marks of \( x_{kn}' \) and \( a_k' \) for notational ease.
The optimal \( \hat{s}_{kn} \) is the function of the amplitudes \( a_k \) and TDS \( \tau \). Now we substitute \( \hat{s}_{kn} \) to (6) and find ML estimates of \( a_k \) and \( \tau \):
\[
\arg\min_{a_k, \tau} \frac{1}{T} \sum_{k=-T/2+1}^{T/2} J_k^{(1)}(a_k, \tau)
\]
\[
\text{s.t. } a_{mk} \geq 0, a_{rk} = 1, \tau_{rr} = 0 \forall m, k, \quad (14)
\]
\[
J_k^{(1)}(a_k, \tau) = \mathbb{E}\left[ \left\| x_{kn} - \frac{g_k^H(a, \tau)x_{kn}}{g_k^H(a, \tau)g_k(a, \tau)}g_k(a, \tau) \right\|^2 \right],
\]
\[
(15)
\]
where \( \mathbb{E}[\cdot] \) denotes the expectation operator, which is replaced by the time average in practice by assuming the ergodic process. By expanding (15) and ignoring constant terms that include neither \( a_k \) nor \( \tau \), the minimization of the sum of \( J_k^{(1)}(a_k, \tau) \) is reduced to the maximization of the following objective function \( J(a_k, \tau) \):
\[
\arg\max_{a_k, \tau} J(a_k, \tau) \quad \text{s.t. } a_{mk} \geq 0, \tau_{rr} = 0 \forall m, k, \quad (16)
\]
\[
J(a_k, \tau) = \frac{1}{T} \sum_{k=-T/2+1}^{T/2} J_k(a_k, \tau),
\]
\[
J_k(a_k, \tau) = \frac{g_k^H(a_k, \tau)V_k g_k(a_k, \tau)}{a_k \tilde{a}_k},
\]
\[
(17)
\]
\[
V_k = \mathbb{E}\left[ x_{kn} x_{kn}^H \right],
\]
\[
(19)
\]
where we use the relationship of \( g_k^H(a_k, \tau)g_k(a_k, \tau) = a_k^2 \tilde{a}_k \), computed form (3), for the denominator. Since the objective function (17) consists of the CC functions for every pair of sensors (which comes from the covariance matrix (19)), we call this objective function MCC. Additionally, the MCC is invariant to the scale of the amplitudes \( a_k \), i.e., \( J_k(a_k, \tau) = J_k(\gamma a_k, \tau) \) with any real numbers \( \gamma \). Hence, the constraint \( a_{rk} = 1 \) (actually, \( a_{rk} = \sigma_r \) due to (13)) can be satisfied by the normalization at the end of the optimization sequences. Finally, our goal is to estimate the subsample TDS \( \tau \) that maximize the MCC (17). We also estimate amplitudes \( a_k \), which may contribute to improving the accuracy of TDE.

In section IV, we first estimate \( \tau \) in the case of \( M = 2 \) as the simplest scenario (i.e., we estimate only one TD), where we introduce the theories essential for solving this optimization problem and obtaining the highly accurate subsample TD estimate. In section V, we then generalize the algorithm described in section IV in the case of \( M \) sensors.

IV. SUBSAMPLE TIME DELAY ESTIMATION

In this section, we explain the technique of maximizing a GCC function by the auxiliary function method as the special case of the proposed method, which was originally presented in the conference paper [38].

A. Problem Formulation

In this section, we consider estimating a subsample TD between two observed signals. For simplicity, we here assume that the amplitude \( a_{mk} \) is 1 for all elements and the variances \( \sigma_m^2 \) are common for all sensors. In this case, the RTF (3) is
\[
g_k(\tau_2) = [1, e^{-j\omega_k \tau_{21}}]^T,
\]
where the reference sensor is set to \( r = 1 \). For notational ease, we hereafter use \( \tau = \tau_{21} \) in this section. Then, the objective function (18) at the \( k \)th frequency becomes
\[
J_k(\tau) = V_{12k} e^{-j\omega_k \tau} + V_{21k} e^{j\omega_k \tau},
\]
where \( V_{ijk} \) is the \((i, j)\) element of the Hermitian matrix \( V_k \), and we omit the constant terms that do not include \( \tau \). Then, by using the conjugate relationship of the first and second terms, we obtain the following optimization problem (16):
\[
\arg\max_{\tau} J(\tau),
\]
\[
J(\tau) = \frac{2}{T} \sum_{k=-T/2+1}^{T/2} V_{21k} e^{j\omega_k \tau},
\]
Equation (23) exactly means the inverse discrete Fourier transform (DFT) of the cross spectrum, i.e., the CC function between the real discrete signals observed by the sensors, \( x_{1kn} \) and \( x_{2kn} \), in the time domain. From the above, the optimization problem (16) reduces to the maximization of the CC function. Note that the relevance between the maximization of the CC function and estimation of the TD on the basis of ML estimation was also discussed in [36].

Now, we consider finding a continuous variable \( \tau \in \mathbb{R} \) maximizing the continuous function (23). We can rewrite (23) as a sum of cosine functions using the conjugate symmetry of \( V_{21k} \) and Euler’s formula,
\[
J(\tau) = \frac{1}{T} \sum_{k=0}^{T/2} A_k \cos(\omega_k \tau + \phi_k),
\]
where \( A_k = 2\beta_k |V_{21k}|, \phi_k = \angle V_{21k}, \angle \) takes the phase of input argument, and \( \beta_0 = \beta_{T/2} = 1 \) and \( \beta_k = 2 \) for \( k \notin \{0, T/2\} \). The arbitrary weight function \( W_k \) can be introduced by replacing the definition of \( A_k \) with \( A_k = 2\beta_k |W_k V_{21k}| \). In this case, the objective function corresponds to the GCC function. Our goal is thus to compute the TD estimate with subsample precision by maximizing (24).

B. Auxiliary Function Method for Subsample TDE

The auxiliary function method (also known as the majorization–minimization (MM) algorithm [41], [42]) is well known as the generalization of the expectation–maximization algorithm and applied to various algorithms [43]–[46]. For adapting it to our problem, we would like to find an auxiliary function \( Q_s(\tau, \theta) \) such that
\[
J(\tau) \geq Q_s(\tau, \theta) \quad \text{for any } \tau \text{ and } \theta,
\]
\[
\text{for any } \tau, \exists \theta = f(\tau) \text{ such that } J(\tau) = Q_s(\tau, \theta),
\]
where \( \theta = (\theta_0, \theta_1, \ldots, \theta_{T/2}) \) are auxiliary variables. Provided such a \( Q_k(\tau, \theta) \) exists and given an initial estimate \( \hat{\theta}^{(0)} \), the following sequence of updates is guaranteed to converge to a local maximum:
\[
\theta^{(\ell)} = f(\hat{\theta}^{(\ell)}), \quad \hat{\theta}^{(\ell+1)} = \arg\max_{\tau \in \mathbb{R}} Q_s(\tau, \theta^{(\ell)}),
\]
where \( \ell \) is the iteration index.

C. Quadratic Auxiliary Function for Continuous GCC

This section provides a quadratic auxiliary function for $J(\tau)$, i.e., $J(\tau) \geq Q_s(\tau, \theta)$.

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

$$Q_s(\tau, \theta) = \frac{1}{T} \sum_{k=0}^{T/2} \left\{ -\frac{A_k}{\theta_k} \sin \theta_k (\omega_k \tau + \phi_k + 2\nu_k \pi)^2 + \cos \theta_k + \frac{1}{2} \theta_k \sin \theta_k \right\},$$

(26)

where $\nu_k \in \mathbb{Z}$ is such that $|\omega_k \tau + \phi_k + 2\nu_k \pi| \leq \pi$. The auxiliary variables are $\theta_k$ and $\nu_k$, then $Q_s(t, \theta) = J(\tau)$ when

$$\theta_k = \omega_k \tau + \phi_k + 2\nu_k \pi.$$  

(27)

This theorem is a direct consequence of the following inequality for a cosine function, which is of general interest.

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

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

(28)

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

Proof of Theorem 1 and Proposition 1 are presented in the appendix. Fig. 2 shows examples of the auxiliary functions for a cosine function (28) and Fig. 3 is for the objective function (24).

D. Derivation of Auxiliary Function and Update Rules

Since $Q_s(\tau, \theta)$ is a quadratic function, it is easily maximized with respect to $\tau$ by setting its derivative to zero:

$$\frac{\partial Q_s(\tau, \theta)}{\partial \tau} = -\sum_{k=0}^{T/2} A_k \omega_k \sin \frac{\theta_k}{\theta_k} (\omega_k \tau + \phi_k') = 0,$$

(29)

where $\phi_k' = \phi_k + 2\nu_k \pi$. Therefore, the maximizer is

$$\hat{\tau} = \max_{\nu_k} \left\{ -\frac{\left( \omega_k \tau + \phi_k + 2\nu_k \pi \right)}{2} \right\}.$$  

(30)

Now, under the condition of equality (27), we can substitute $\phi_k' = \theta_k - \omega_k \tau$ for (30) and obtain the final update rules:

$$\nu_k^{(\ell)} \leftarrow \text{round}\left\{ -\frac{\left( \omega_k \tau^{(\ell)} + \phi_k \right)}{2\pi} \right\},$$

(31)

$$\theta_k^{(\ell)} \leftarrow \omega_k \tau^{(\ell)} + \phi_k + 2\nu_k^{(\ell)} \pi, \quad k = 0, \ldots, \frac{T}{2},$$

(32)

$$\tau^{(\ell+1)} \leftarrow \tau^{(\ell)} - \frac{\sum_{k=0}^{T/2} A_k \omega_k^2 \left( \sin \theta_k^{(\ell)} / \theta_k^{(\ell)} \right) \phi_k^{(\ell)}}{\sum_{k=0}^{T/2} A_k \omega_k^2 \left( \sin \theta_k^{(\ell)} / \theta_k^{(\ell)} \right)}.$$

(33)

where round[·] is a function that rounds the input to the nearest integer. Interestingly, the second term of (33) is a weighted sum of the auxiliary variables scaled by the normalized angular frequency, i.e., $\theta_k^{(\ell)}/\omega_k$, and $\theta_k^{(\ell)}/\omega_k$ corresponds to the candidate of the TD estimate at the frequency $k$.

V. AuxTDE: Consistent Time Delay Estimation

In this section, we propose AuxTDE, the method for estimating multiple sub-sample time delays (STDs) simultaneously. This is the generalization of the method introduced in the previous section.

A. Technical Approach for TDE

As we mentioned in section III, we aim to maximize the objective function (17) to obtain subsample TD estimates. To solve the joint optimization problem (16), we first consider to optimize $\tau$ with fixed $a_k$. Then, the denominator of (17) is constant that does not include $\tau$. Now, the question is how to
maximize the sum of numerator with respect to \( \tau \), which has the following structure:

\[
g_k^H(\tau) V_k g_k(\tau) = \begin{pmatrix} a_1 e^{-i \omega_k \tau_1} \\ a_2 e^{-i \omega_k \tau_2} \\ \vdots \\ a_M e^{-i \omega_k \tau_M} \end{pmatrix}^H \begin{pmatrix} a_1 e^{-i \omega_k \tau_1} \\ a_2 e^{-i \omega_k \tau_2} \\ \vdots \\ a_M e^{-i \omega_k \tau_M} \end{pmatrix},
\]

where the parameters to be optimized are the exponent. Unfortunately, the maximization of the above function has no closed-form solution. Therefore, we consider applying the auxiliary function method as in section IV.

The objective function at the \( k \)th frequency (18) can be rewritten as a sum of the cosine function using the conjugate symmetry of \( V_k \) and Euler’s formula, the same as the derivation of (24),

\[
J_k(\tau) = \sum_{i=1}^{M} \sum_{j=1}^{M} A_{ijk} \cos(\omega_k \tau_{ij} + \phi_{ijk}),
\]

where \( A_{ijk} = \delta_k \delta_{ijk} |V_{ijk}|, \phi_{ijk} = \angle V_{ijk}, \) and the denominator is omitted. It is worth noting that because the difference between two TDs, i.e., \( \tau_{ij} = \tau_j - \tau_i \) in (35), is equal to \( t_j - t_i \), the objective function is independent of the reference sensor \( r \). Now, we propose the quadratic auxiliary function for MCC, which has the closed-form solution. Here, we consider updating \( \tau, a_k, \) and \( \sigma_m \) alternately.

B. Quadratic Form Auxiliary Function for TDE

Similarly to subsection IV-B, we design the auxiliary function \( Q(\tau, \theta) \) for multiple TDs \( \tau \) that satisfies the following properties:

- \( J(\tau) \geq Q(\tau, \theta) \) for any \( \tau \) and \( \theta \).
- For any \( \tau, \exists \theta = f(\hat{\tau}) \) such that \( J(\tau) = Q(\hat{\tau}, \theta) \), where \( \theta = (\theta_{ijk})_{1 \leq i,j,k \leq M, 0 \leq k \leq T/2} \) are auxiliary variables. Provided such a \( Q(\tau, \theta) \) exists and given an initial estimate \( \hat{\tau}^{(0)} \), the following sequence of updates is guaranteed to converge to a local maximum:

\[
\theta^{(\ell)} = f(\hat{\tau}^{(\ell)}), \quad \hat{\tau}^{(\ell+1)} = \arg \max_{\tau \in \mathbb{R}} Q(\hat{\tau}^{(\ell)}, \theta^{(\ell)}).
\]

We here propose an auxiliary function for \( J(\tau) \), i.e., \( J(\tau) \geq Q(\tau, \theta) \).

Theorem 2. The following is an auxiliary function for \( J(\tau) \):

\[
Q_k(\tau, \theta) = \frac{1}{T} \sum_{k=0}^{T/2} Q_k(\tau, \theta),
\]

\[
Q_k(\tau, \theta) = \sum_{i=1}^{M} \sum_{j=1}^{M} \left\{ - B_{ijk} (\omega_k \tau_{ij} + \phi_{ijk} + 2 \nu_{ijk} \pi)^2 + \cos \theta_{ijk} + \frac{1}{2} \theta_{ijk} \sin \theta_{ijk} \right\},
\]

\[
B_{ijk} = \frac{A_{ijk} \sin \theta_{ijk}}{2 \theta_{ijk}},
\]

where \( \nu_{ijk} \in \mathbb{Z} \) is such that \( |\omega_k \tau_{ij} + \phi_{ijk} + 2 \nu_{ijk} \pi| \leq \pi \). The auxiliary variables are \( \theta_{ijk} \) and \( \nu_{ijk} \) and \( Q(\tau, \theta) = J(\tau) \) holds when

\[
\theta_{ijk} = \omega_k \tau_{ij} + \phi_{ijk} + 2 \nu_{ijk} \pi, \quad \forall i, j, k.
\]

This theorem is a direct consequence of Proposition 1 with regards to a cosine function. The auxiliary function (38) can be rewritten as the vector quadratic form:

\[
Q_k(\tau, \theta) = -\tau^T C_k \tau - 2 c_k^T \tau + \text{Const.},
\]

\[
C_k = \omega_k^2 \{ \text{diag}(b_k) - B_k \},
\]

\[
b_k = \left( \sum_{i=1}^{M} B_{ik1}, \ldots, \sum_{i=1}^{M} B_{ikM} \right)^T,
\]

\[
c_k = \omega_k \left( \sum_{i=1}^{M} B_{ik1} \phi_{1ik}, \ldots, \sum_{i=1}^{M} B_{ikM} \phi_{Mik} \right)^T,
\]

where \( B_k = (B_{ij})_{1 \leq i,j \leq M} \) and \( \phi_{ijk} = \phi_{ijk} + 2 \nu_{ijk} \pi \). \( C_k \) is the positive semidefinite matrix [47] and thus \( -\tau^T C_k \tau \) is the convex upward function. We show an example of the objective function and auxiliary functions in Figs. 4(a)–(d), where \( M = 3 \) (that is, there are two TDs). The objective function, the MCC, is globally lower bounded by the proposed auxiliary function \( Q(\tau, \theta) \) at any point and has exactly one point of tangency.

C. Derivation of Auxiliary Function and Update Rules

To maximize the auxiliary function \( Q(\tau, \theta) \) under the constraint of \( \tau_{rr} = 0 \), we find the stationary point of the function \( \sum_{k=0}^{T/2} Q_k(\tau, \theta) - \lambda \tau_{rr} \), where \( \lambda \) is the Lagrange
multiplier. Since \( Q_k(\tau, \theta) \) is a quadratic form, it is easily maximized with respect to \( \tau \) by setting its derivative to zero:

\[
\frac{\partial Q(\tau, \theta)}{\partial \tau} = -\sum_{k=0}^{T/2} \{2C_k \tau + 2c_k\} - \lambda e_r = 0,
\]

where \( e_r \) is a unit vector whose only \( r \)th element is 1. Then, we obtain

\[
\bar{\tau} = -\left( \sum_{k=0}^{T/2} C_k \right)^{-1} \sum_{k=0}^{T/2} \bar{c}_k,
\]

where \( C_k \) is a \( M - 1 \times M - 1 \) matrix eliminating the \( r \)th row and column from \( C_k \), namely, \( C_k = (C_{ij})_{k,1 \leq i,j \leq M,i,j \neq r} \). Similarly, \( \bar{\tau} = (\bar{\tau}_r)_{1 \leq r \leq M,i,j \neq r} \) and \( \bar{c}_k = (c_k')_{1 \leq k \leq M,i,j \neq r} \). Finally, under the condition of equality (40), we can substitute \( \theta_{ijk} = \theta_{ijk} - \omega_k \tau_{ij} \) and obtain the update rules:

\[
\begin{align*}
\nu^{(t)}_{ijk} &\leftarrow \text{round}\left(-\left(\omega_k \tau_{ij}^{(t)} + \phi_{ijk}\right)/2\pi\right), \\
\theta^{(t)}_{ijk} &\leftarrow \omega_k \tau_{ij}^{(t)} + \phi_{ijk} + 2\nu^{(t)}_{ijk}, \forall i,j,k,
\end{align*}
\]

\[
\bar{\tau}^{(t+1)} = \bar{\tau}^{(t)} - \left( \sum_{k=0}^{T/2} C_k^{(t)} \right)^{-1} \sum_{k=0}^{T/2} \bar{c}_k^{(t)},
\]

where \( \bar{c}_k^{(t)} = (c_k')_{1 \leq k \leq M,i,j \neq r} \) and

\[
c^{(t)}_{ijk} = \omega_k \sum_{i=1}^{M} B_{ijk} \frac{\theta_{ijk}}{\omega_k}.
\]

Note that the above update rules are in complete agreement with those in the case of two sensors (31)–(33), and hence the TD estimate is also identity.

**D. Amplitude Estimation**

Although our purpose is to estimate subsample TDs, we need to also estimate the unknown parameter \( \alpha_k \). One of the solutions is the use of fixed amplitudes \( \alpha_k = 1 \) for all \( k \). However, we here propose the following method for estimating \( \alpha_k \), which may result in improved TD estimates.

First, we arrange the numerator of \( J_k(\alpha_k, \tau) \) as

\[
g_k^H(\alpha_k, \tau)V_k g_k(\alpha_k, \tau) = a_k^T V_k^H(\tau) a_k,
\]

\[
V_k^H(\tau) = V_k \odot P_k(\tau),
\]

\[
P_k(\tau) = p_k(\tau)p_k^H(\tau),
\]

\[
p_k(\tau) = [e^{j\omega_k \tau_1} \ldots e^{j\omega_k \tau_M}]^T,
\]

where \( V_k^H(\tau) \) is an Hermitian matrix and \( P_k(\tau) \) is an \( M \times M \) matrix\(^2\). At fixed \( \tau \), the optimization problem (16) becomes

\[
\arg \max_{\alpha_k} 1 \sum_{k=0}^{T/2} \beta_k J_k(\alpha_k) \quad \text{s.t.} \quad a_{mk} \geq 0 \quad \forall m,k,
\]

\[
J_k(\alpha_k) = \frac{a_k^T V_k^H(\tau) a_k}{a_k^T a_k} = \frac{\Re[V_k^H(\tau) a_k]}{a_k^T a_k},
\]

where \( \Re[\cdot] \) takes the real part of the input argument, and again \( \beta_0 = \beta_{T/2} = 1 \) and \( \beta_k = 2 \) for \( k \notin \{0, T/2\} \). The objective function (56) is known as the Rayleigh quotient, where \( a_k \) is the non-zero (and also non-negative in our problem) vector. Hence, the optimization problem is the maximization of the Rayleigh quotient with a non-negative constraint, which may be of general interest.

1) **Unconstrained case**: The solution of this maximization problem without any constraint is given by the eigenvalue decomposition (e.g., [42]). Since the objective function (56) is invariant to the scale of the amplitudes, we optimize \( \alpha_k \) under the constraint \( \|a_k\|_2^2 = 1 \) and compensate for \( \alpha_k \) to 1 at the end of update sequences. Then, using the method of Lagrange multipliers, we find the stationary point by taking the gradient with respect to \( \alpha_k \) and setting it to zero as

\[
\frac{\partial J_k(\alpha_k)}{\partial a_k} = V_k^H a_k - \lambda_k a_k = 0,
\]

where \( \lambda_k \) is the \( k \)th Lagrange multiplier.

This result is well known as the eigenvalue problem. Although any eigenvector of \( V_k^H \) used as \( \alpha_k \) satisfies (57), we choose the eigenvector corresponding to the largest eigenvalue (hereafter, we simply denote it as the largest eigenvector) for maximizing the objective function (56).

2) **Constrained case**: One solution for the maximization of the Rayleigh quotient with non-negative constraint is the projection. Since the largest eigenvector can take a negative value, we project its elements \( \bar{a}_{mk} \) to the positive domain

\[
\bar{\alpha}_{mk} \leftarrow \max(\bar{\alpha}_{mk}, 0) \quad \forall m,
\]

to satisfy the non-negative constraint.

Although the largest eigenvector is absolutely the solution of (56), there is no guarantee that the projected eigenvector maximizes the objective function. Thus we propose the alternative method based on the auxiliary function method.

**Theorem 3.** Let \( \|a_k\|_2^2 = 1 \) for all \( k \). The following is an auxiliary function for the Rayleigh quotient \( J_k(\alpha_k) \) [42],

\[
Q_k(\alpha_k, \theta) = 2\theta_k V_k^H a_k - \theta_k V_k^H \theta_k,
\]

where \( \theta = [\theta_0 \ldots \theta_{T/2}]^T \) are auxiliary variables and \( Q_k(\alpha_k, \theta) = J_k(\alpha_k) \) holds when

\[
\theta_k = a_k.
\]

The auxiliary function (59) is the linear form of the vector \( \alpha_k \) and is easily maximized even under the non-negative constraint. Then, the update rules are

\[
\begin{align*}
a_k^{(t+1)} &\leftarrow V_k^H a_k^{(t)}, \\
a_{mk}^{(t+1)} &\leftarrow \max(a_{mk}^{(t+1)}, 0) \quad \forall m,
\end{align*}
\]

\[
a_k^{(t+1)} \leftarrow a_k^{(t+1)} / \|a_k^{(t+1)}\|_2.
\]

These update sequences are guaranteed to converge to a local maximum, whereas the projected largest eigenvector (58) does not attain it unless all elements are positive without projection.

Interestingly, there is a well-known algorithm, namely, the power method (power iteration), for estimating the largest eigenvector. The power method iteratively updates the eigenvector estimate in (61) and (63). The above algorithm shows that we can obtain the local maximum in the same scheme

---

\(^2\)Properties of these matrices are discussed in the appendix.
even under the non-negative constraint by applying the projection to the positive domain (62).

3) Shared amplitude case: We can consider the observation model with a shared amplitude vector for all frequencies, that is, $a_k$ is identical for all $k$. The amplitude estimation with this alternative model can be easily realized by using averaged $V'_k$:

$$V'_{\text{mean}} = \frac{1}{T/2 + 1} \sum_{k=0}^{T/2} V'_k,$$

in update sequences (61)–(63). We expect that this alternative model is robust against the estimation error in amplitudes and measurement environments.

E. Update of Noise Variances

To compute the observations and RTF weighted by the standard deviation $\sigma_m$ defined in (12) and (13), respectively, we finally update the variance $\Sigma$ of the complex multivariate Gaussian distribution $N_c(\mu, \Sigma)$ as follows:

$$\sigma^2_m \leftarrow \frac{1}{N_T} \sum_{k=0}^{T/2-1} \sum_{n=0}^{N-1} \beta_k |x_{kn} - \hat{s}_{kn} g_{mk}(a_m, \tau_m)|^2,$$

where $g_{nk}(a_m, \tau_m) = a_m e^{-i\omega_m \tau_m}$ and $\hat{s}_{kn}$ is computed by (11).

F. Algorithm of AuxTDE

Finally, we summarize the update sequences of the AuxTDE in Algorithm 1, where three types of iteration exist, for the update of $\tau$, $a_k$, and their alternate updates indexed by $\ell$, $\ell_a$, and $\ell_e$, respectively. The maximum iterations $L_{\ell}$, $L_{\ell_a}$, and $L_{\ell_e}$ can be 1. The initialization of $\tau$ can be performed by the pairwise method. For instance, the discrete maximum (GCC method [22]) and the result of parabolic interpolation [28] can be used. Basically, better initial estimates lead to faster convergence, and parabolic interpolation is thus better in practice.

VI. EXPERIMENTAL ANALYSIS OF THE AUXTDE ALGORITHM

A. Empirical Convergence to the Local Maximum

First, we confirm the convergence of the AuxTDE by depicting the objective function, where we simulated $M = 2$ observations whose sub-sample TD was set to 2.0996 samples. We used English female speech as a target signal sampled at 16 kHz and added white Gaussian noise to each microphone with signal-to-noise ratio (SNR) of 10 dB. Fig. 5 shows the convergence of the objective function (17) with the AuxTDE for different initial values. We set the initial value of the proposed method to every three samples from 2 to 29. In accordance with this figure, we can confirm that the proposed method will converge to the global maximum if an appropriate initial value is given. Moreover, this figure shows the guaranteed monotonic increase in the objective function for any initial value. The initial value must be picked from the unimodal period, including the global maximum, to reach it, where the range is between −20 to 22 samples in this figure. Basically, the conventional GCC method is a good way to obtain such an initial estimate. Even when the initial estimate is outside the appropriate range, the convergence to the local maximum is always guaranteed owing to the characteristic of the auxiliary function method. Additionally, the better the initial estimate is (e.g., in the case of using the parabolic interpolation), the faster the convergence is.
B. Empirical Convergence and Consistent TD Estimates

As we mentioned in section II-C, the AuxTDE attains consistent TD estimates owing to the observation model. Although we must set the reference sensor for computing TDs since these are relative values, the AuxTDE can obtain the same TD estimates regardless of the reference sensor. This property can be confirmed by verifying the objective function. Here, we used pyroomacoustics [48] to simulate a reverberant environment and generated eight microphone signals with a target source, as shown in Fig. 6. In this experiment, we focused on the TDE algorithm of the AuxTDE and fixed the amplitude $a_k$ and the variance $\sigma^2_m$ to 1 for all elements. We set initial values of $\tau$ as the estimates by the pairwise method using the parabolic interpolation [28] of the GCC method [22] whose weights were one for all frequencies. For the pairwise method, there were eight choices of the reference microphone $r$; thus, we computed the same number of the TD estimate vector $\tilde{\tau}_r$. Then, we computed the objective function of the AuxTDE (17) with every $\tilde{\tau}_r$ and showed them in Fig. 7 at the 0th iteration.

The objective function values at the 0th iteration in Fig. 7 show the dependence of the pairwise GCC method on the reference microphone $r$. This result implies that the performance of TD estimation also varied. Interestingly, the best performance in terms of the objective function was achieved with $r = 3$, whereas the worst one was achieved with $r = 4$ (see Fig. 6). It is difficult to predict the best reference microphone in advance, even when the actual layout of the source and microphones was known. Moreover, since we cannot obtain true TD values in practice, the evaluation criterion for the pairwise method is also unclear.

In contrast, the AuxTDE with these initial estimates monotonically increased the objective function and converged to the same local maximum regardless of the reference microphone, as shown in Fig. 7. This result indicates that consistent TDs were obtained owing to the joint optimization (16) and the transfer system model (3). Additionally, it can be said that these results are the best in terms of ML. Similar to the result in subsection VI-A, the better initial estimate led to faster convergence.

VII. EXPERIMENTS OF TIME DELAY ESTIMATION

A. Experimental Condition

In this section, we evaluated the performance of TD estimation. We used the pyroomacoustics [48] to simulate reverberant room environments. We synthesized $M$ observed signals sampled at 16 kHz with simulated room impulse responses with a reverberation time of approximately 200 ms. The target signal of 5 s was randomly generated following normal Gaussian distribution and was contaminated by additive Gaussian noise, where SNR was set to 20 dB. The target source and microphones are randomly located in a room of $3 \times 4 \times 3$ m size, as in the example shown in Fig. 6. We tested three types of microphone alignment: widely placed $M = 4$ and $M = 8$ microphones assuming distributed microphone arrays (DMAs), where they are located at least 0.2 m away from each other, and closely placed $M = 8$ microphones assuming an ordinary microphone array. Note that we assumed that all microphones are synchronized, and no sampling frequency mismatch problem [20] occurred. We performed STFT with a rectangle window for the observed signals, where the window length is 4096 samples, and each frame is half-overlapped.

We evaluated three types of AuxTDE algorithm for amplitude estimation: the original AuxTDE, denoted as AuxTDE_freqAmp, estimates both TDs and frequency-dependent amplitudes simultaneously as shown in Algorithm 1, AuxTDE_shrdAmp estimates shared (frequency-independent) amplitude by the algorithm described in subsection V-D3, and AuxTDE_unitAmp estimates only TDs with fixed amplitudes ($a_k = 1 \ \forall k$). The number of iterations of each method is listed in Table I, and all the methods update the variances.

For comparison, we evaluated three types of pairwise methods: the GCC method [22] (PW-GCC), the GCC method with parabolic interpolation [28] (PW-Parafit), and the GCC method with the AuxTDE for two channels (PW-AuxTDE). Although
the AuxTDE is applicable with three or more microphones, we used it here as the pairwise method merely for comparison. We used every microphone as the reference one for all the methods and obtained M TD estimate vectors \( \hat{\tau}_r \).

For the evaluation criterion, we used root mean square errors (RMSEs) between the estimated TDs \( \hat{\tau}_{r,p} \) and computed ones \( \tau_{r,p} \) defined as:

\[
RMSE = \sqrt{\frac{1}{PM(M-1)} \sum_{p=1}^{P} \sum_{r=1}^{M} \left( \hat{\tau}_{r,p} - \tau_{r,p} \right)^2},
\]

where \( P \) denotes the number of simulations, and the subscript \( p \) denotes the simulation index. \( P \) was 1200 in this experiment. Since the true TDs were unknown, we computed the TDs from the distance between the target source and each microphone and used it as the ground truth instead, where the speed of sound was 343 m/s (pyroacoustics default). Note that we eliminated several gross error cases from the evaluation. The occurrence of gross errors depended on the conditions (e.g., microphone positions and initial estimates) and was approximately 0.1\% in this experiment.

In addition to the evaluation of RMSE, we evaluate how inconsistent the TD estimates are. On the basis of Definition 1 for the consistent TDs, we define the mean absolute inconsistent delay (MID) as follows:\(^3\):

\[
MID = \left( \frac{1}{M-1} \right)^2 \sum_{r=1}^{M} \sum_{m=1}^{M} |\tau_{r,m} - (\tau_{r,\hat{r}} + \tau_{\hat{r},m})|, \quad (67)
\]

where \( r = 1 \) in this experiment. Clearly, MID is 0 if the TD estimates are completely consistent; otherwise, it takes a high value. To evaluate the MID, we thus need to perform TD estimation \( M \) times in total by setting each sensor as the reference one.

\(^3\)The TD estimate at the reference microphone \( \tau_{r,\hat{r}} \) is always zero, and the number of TD estimates is thus \( M - 1 \).

\(^4\)When \( r = \hat{r}, |\tau_{r,m} - (\tau_{r,\hat{r}} + \tau_{\hat{r},m})| \) is always zero for all \( m \), and we thus eliminated this case from the parameter.

### Table I
**Number of iterations of each AuxTDE variant.**

| Method              | \( \ell_r \) | \( \ell_a \) | \( \ell_c \) |
|---------------------|--------------|--------------|--------------|
| AuxTDE\_freqAmp     | 10           | 10           | 3            |
| AuxTDE\_unitAmp     | 10           | -            | 3            |
| AuxTDE\_shrdAmp     | 10           | 10           | 3            |

### Table II
**RMSEs between the estimated and ground truth TDs for each method and environment.**

| Method          | \# of microphones | 4 (DMA) | 8 (DMA) | 8 (Array) |
|-----------------|-------------------|--------|--------|----------|
| PW-GCC          | 16.5 \times 10^{-2} | 21.7 \times 10^{-2} | 21.4 \times 10^{-2} |
| PW-Parafit      | 6.68 \times 10^{-2} | 8.71 \times 10^{-2} | 8.83 \times 10^{-2} |
| PW-AuxTDE       | 1.09 \times 10^{-2} | 1.40 \times 10^{-2} | 2.22 \times 10^{-2} |
| AuxTDE\_unitAmp | 3.88 \times 10^{-12} | 3.88 \times 10^{-12} | 3.73 \times 10^{-12} |
| AuxTDE\_freqAmp | 2.85 \times 10^{-12} | 0.941 \times 10^{-12} | 7.79 \times 10^{-12} |
| AuxTDE\_shrdAmp | 0.162 \times 10^{-12} | 0.122 \times 10^{-12} | 0.290 \times 10^{-12} |

### Table III
**MIDs for each method and environment.**

| Method          | \# of microphones | 4 (DMA) | 8 (DMA) | 8 (Array) |
|-----------------|-------------------|--------|--------|----------|
| PW-GCC          | 16.5 \times 10^{-2} | 21.7 \times 10^{-2} | 21.4 \times 10^{-2} |
| PW-Parafit      | 6.68 \times 10^{-2} | 8.71 \times 10^{-2} | 8.83 \times 10^{-2} |
| PW-AuxTDE       | 1.09 \times 10^{-2} | 1.40 \times 10^{-2} | 2.22 \times 10^{-2} |
| AuxTDE\_unitAmp | 3.88 \times 10^{-12} | 3.88 \times 10^{-12} | 3.73 \times 10^{-12} |
| AuxTDE\_freqAmp | 2.85 \times 10^{-12} | 0.941 \times 10^{-12} | 7.79 \times 10^{-12} |
| AuxTDE\_shrdAmp | 0.162 \times 10^{-12} | 0.122 \times 10^{-12} | 0.290 \times 10^{-12} |

### B. Results and Discussion

Table II shows the RMSEs, the results of TDs estimation. The theoretical error in PW-GCC is 0.25, and values close to this error were obtained. PW-Parafit significantly improved the estimation accuracy with quite low computational cost, and PW-AuxTDE attained greater improvement with the efficient iterative algorithm. Their RMSEs were almost the same for all the microphone alignments.

The RMSEs of AuxTDE algorithms except for PW-AuxTDE were superior to that of the pairwise methods in every case. Additionally, the performance was improved by using eight microphones than four microphones. Since our model includes spatial information contained in the entire observation, AuxTDE could utilize the consistency in TD parameters. Furthermore, the performance with DMAs was relatively higher than in the case of using an array of closely placed microphones in this experiment. This result implies that one of the suitable applications of the AuxTDE is a DMA (and other distributed sensor systems), which consists of widely placed sensors and has broad spatial information.

The simultaneous estimation of the amplitude (AuxTDE\_freqAmp) degraded the performance in most cases compared with AuxTDE\_unitAmp even though the performance of AuxTDE\_freqAmp was superior to that of the pairwise methods. As one reason, we can consider that the constant amplitude used in AuxTDE\_unitAmp (i.e., \( a_k = 1 \) for all \( k \)) was an excellent a priori for TD estimation. Moreover, it is possible that estimating frequency-dependent amplitudes overfitted the acoustic environment. Estimating shared amplitudes (AuxTDE\_shrdAmp) may be the better solution for some situations such as when using DMA. For example, when the gain of each sensor differs, the mechanism of AuxTDE\_shrdAmp may be able to reduce the negative effect due to their gain differences.

Table III shows the MID of the TD estimates. The MIDs of the pairwise methods were considerably high, and the order of MIDs was the same as the RMSEs (see also Table II). This means that the TD estimates were inconsistent. Therefore, there should be one best microphone that should be used as the reference microphone; however, the method to find it is unclear. On the other hand, the proposed methods that estimate all TDs simultaneously showed markedly low MIDs regardless of the microphone alignment. This means that the TD estimates were independent of the reference microphone. From the above, we can confirm the effectiveness of the
proposed AuxTDE for TD estimation.

C. TD Estimation with a Number of Microphones

Finally, we investigated the relationship between the accuracy in TDE and the number of microphones to compare the proposed methods and the pairwise methods further. Experimental conditions are identical to those described in subsection VII-A except for the number of microphones $M$. It varies from 2 to 16, and they are located at least 0.2 m away from each other, assuming DMAs.

Figs. 8(a) and (b) show the RMSE and MID of each method as functions of the number of microphones, respectively. Note that the results of PW-GCC and PW-Parafit are omitted because their performance changes with respect to $M$ tended to be the same as that of PW-AuxTDE. The TD estimates of each method except for AuxTDE_freqAmp are theoretically identical when $M = 2$.

From Fig. 8(a), the RMSE of PW-AuxTDE was independent of the number of microphones $M$. Since the pairwise method only uses partial spatial cue between two selected sensors, there are no benefits of increasing the number of sensors. In contrast, the MID increased (i.e., worsened) with increasing $M$, as shown in Fig. 8(b). The MID is theoretically zero when $M = 2$ since the interchange of the reference sensor corresponds to the time reversal of the CC function. The choices of reference sensors increased with increasing $M$, and as a result, the MID became high.

In contrast to the above results, AuxTDE algorithms except for PW-AuxTDE improved the performance of TDE with increasing number of microphones. This result demonstrated the importance and efficacy of using entire spatial information captured by the microphones and consistent constraint for TDs. Additionally, the MIDs were considerably low (approximately $10^{-12}$); in other words, the TDs estimated by the proposed methods were consistent regardless of $M$. Finally, we concluded that AuxTDE algorithms are effective for TDE, which attains highly accurate and consistent TD estimates.

VIII. CONCLUSIONS

In this paper, we proposed AuxTDE, a novel method for TD estimation, using the auxiliary function method. The joint optimization problem for consistent TDs and amplitudes was considered on the basis of ML estimation. The objective function, MCC function, encodes the CC function of all sensor pairs, which can thus be considered as the multidimensional extension of the CC function. The MCC, which is the non-convex function, was lower-bounded by the quadratic auxiliary function, and efficient update rules that iteratively maximize the MCC were derived. In experiments, we demonstrated important properties of AuxTDE: monotonic increases in the objective function, convergence to the local maximum, and independence against the reference sensor. Additionally, we confirmed the efficacy of the AuxTDE through the experiment of TD estimation, where the AuxTDE attained highly accurate and consistent TD estimates. The future work includes the online extension of the AuxTDE and the heuristic extension of its algorithm.

APPENDIX A

PROOF OF PROPOSITION 1 AND THEOREMS 1 AND 2

A. Proof of Proposition 1

Proof: Let

$$f(\theta) = \cos \theta + \frac{1}{2} \frac{\sin \theta}{\theta_0} \theta^2 - \left( \cos \theta_0 + \frac{1}{2} \theta_0 \sin \theta_0 \right).$$

Then, we have

$$f'(\theta) = -\sin \theta + \frac{\sin \theta_0}{\theta_0} \theta = -\theta \left( \frac{\sin \theta}{\theta} - \frac{\sin \theta_0}{\theta_0} \right).$$
We separately consider the following three cases.

**Case 1:** $0 < |\theta_0| < \pi$

Because $\sin \theta / \theta$ is monotonically decreasing in $0 \leq \theta \leq \pi$, we have
\[
f'(\theta) = \begin{cases} < 0 & (0 \leq \theta < |\theta_0|), \\ = 0 & (\theta = |\theta_0|), \\ > 0 & (|\theta_0| < \theta \leq \pi). \end{cases}
\]

(70)

It thus appears that $f(\theta)$ attains its minimum at $|\theta_0|$. Moreover, $f(\theta_0) = 0$ and thus $f(\theta) \geq 0$ in $0 \leq \theta \leq \pi$. Since $f(\theta)$ is an even function, its minimum value in $-\pi \leq \theta \leq \pi$ is also 0. Because $\cos \theta$ is periodic but $-\theta^2$ is not, $f(\theta + 2\nu \pi) > f(\theta)$ for any $-\pi < \theta < \pi$ and integer $\nu \neq 0$. Therefore, $f(\theta) \geq 0$, with equality if and only if $|\theta| = |\theta_0|$. 

**Case 2:** $\theta_0 = 0$

In this case, for $0 \leq \theta < \pi$, we have
\[
f'(\theta) = \begin{cases} = 0 & (\theta = |\theta_0| = 0), \\ > 0 & (|\theta_0| < \theta \leq \pi), \end{cases}
\]

which means $f(\theta)$ takes a minimum value at $f(0) = 0$ in $-\pi \leq \theta \leq \pi$. Similarly to case 1, we obtain $f(\theta) \geq 0$, with equality if and only if $\theta = 0$. 

**Case 3:** $0 = \theta_0$ or $\theta_0 = -\pi$

In this case, $f(\theta) = \cos \theta + 1$. Therefore $f(\theta) \geq 0$, with equality if and only if $\theta = (2\nu + 1)\pi$ for any $\nu \in \mathbb{Z}$. 

**B. Proof of Theorems 1 and 2**

**Proof:** Because $\cos(\omega_k \tau_{ij} + \phi_{ij}) = \cos(\omega_k \tau_{ij} + \phi_{ij} + 2\nu_{ij} \pi)$ with $\nu_{ij} \in \mathbb{Z}$, and because $\beta_{ijk} a_{ik} a_{jk} A_{ijk} \geq 0$, we can apply Proposition 1 separately to each term of the sum in (24) and (35). 

**APPENDIX B**

**PROPERTIES OF MATRICES IN SUBSECTION V-D**

$P_k$ is a positive semidefinite matrix because the following is satisfied for any complex-valued vector $z$:
\[
z^H P_k z = |p_k^H z|^2 \geq 0.
\]

(72)

In accordance with the Schur product theorem, $V_k'$, which is the Hadamard product of two positive semidefinite matrices, is also the positive semidefinite matrix. Additionally, $\text{rank}(P_k) = 1$ and thus $\text{rank}(V_k') = \text{rank}(V_k)$. When we assume that the signal $s_k$ and noises $u_{kn}$ are uncorrelated, the covariance matrix $V_k$ can be divided into the signal and noise parts as
\[
V_k = \mathbb{E} [x_{kn} x_{kn}^H] = |s_k|^2 g_k g_k^H + \mathbb{E} [u_{kn} u_{kn}^H],
\]

(73)

Following the subadditivity of the rank of the matrix, $\text{rank}(V_k') \leq \text{rank}(\mathbb{E} [u_{kn} u_{kn}^H]) + 1$, where $g_k g_k^H$ is a rank-1 matrix.

**ACKNOWLEDGMENT**

This work was supported by JSPS KAKENHI grant numbers 20H00613 and 19J20420 and JST CREST grant number JPMJCR19A3, Japan. The authors would like to thank Robin Schiebeler for collaboration in the early stage of this work.

**REFERENCES**

[1] J. Chen, J. Benesty, and Y. Huang, “Time delay estimation in room acoustic environments: An overview,” EURASIP J. Adv. Signal Process., vol. 2006, pp. 1–19, Dec. 2006.

[2] M. Brandstein, J. Adcock, and H. Silverman, “A closed-locus location estimator for use with room environment microphone arrays,” IEEE Trans. Speech Audio Process., vol. 5, no. 1, pp. 45–50, 1997.

[3] T. Gustafsson, B. Rao, and M. Trivedi, “Source localization in reverberant environments: modeling and statistical analysis,” IEEE Trans. Speech Audio Process., vol. 11, no. 6, pp. 791–803, 2003.

[4] X. Alameda-Pineda and R. Horaud, “A geometrical approach to sound source localization from time-delay estimates,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 22, no. 6, pp. 1082–1095, 2014.

[5] O. Schwartz and S. Gannot, “Speaker tracking using recursive EM algorithms,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 22, no. 2, pp. 392–402, Feb. 2014.

[6] C. Evers, H. W. Löffmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, “The LOCATA challenge: Acoustic source localization and tracking,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 28, pp. 1620–1643, 2020.

[7] G. Carter, “Time delay estimation for passive sonar signal processing,” IEEE Trans. Acoust., Speech, Signal Process., vol. 29, no. 3, pp. 463–470, Jun. 1981.

[8] P. Protiva, J. Mrkvica, and J. Macháč, “Estimation of wall parameters from time-delay-only through-wall radar measurements,” IEEE Trans. Antennas Propag., vol. 59, no. 11, pp. 4268–4278, Nov. 2011.

[9] Q. Lele, S. Qiang, Y. Tianhong, Z. Lili, and S. Yangpeng, “Time-delay estimation for ground penetrating radar using ESPRIT with improved spatial smoothing technique,” IEEE Geosci. Remote Sens. Lett., vol. 11, no. 8, pp. 1315–1319, Dec. 2014.

[10] J. Capon, “Applications of detection and estimation theory to large array seismology,” Proc. IEEE, vol. 58, no. 5, pp. 760–770, May 1970.

[11] T.-K. Le and N. Ono, “Closed-form and near closed-form solutions for TOA-based joint source and sensor localization,” IEEE Trans. Signal Process., vol. 64, no. 18, pp. 4751–4766, 2016.

[12] T.-K. Le and N. Ono, “Closed-form and near closed-form solutions for TDOA-based joint source and sensor localization,” IEEE Trans. Signal Process., vol. 65, no. 5, pp. 1207–1221, 2017.

[13] Y. Sun, K. C. Ho, and Q. Wan, “Solution and analysis of TDOA localization of a near or distant source in closed form,” IEEE Trans. Signal Process., vol. 67, no. 2, pp. 320–335, 2019.

[14] S. Makino, T.-W. Lee, and H. Sawada, Blind Speech Separation. Berlin: Springer, 2007.

[15] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, “A consolidated perspective on multimicrophone speech enhancement and source separation,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 25, no. 4, pp. 692–730, Jan. 2017.

[16] Y. Wakabayashi, K. Yamaoka, and N. Ono, “Rotation-robust beamforming based on sound field interpolation with regularly circular microphone array,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., pp. 771–775, May 2021.

[17] 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.

[18] A. Chinea, P. Thué, and G. Enzner, “Double-cross-correlation processing for blind sampling-rate and time-offset estimation,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 29, pp. 1881–1896, 2021.

[19] A. D. Coulson, “Maximum likelihood synchronization for OFDM using a pilot symbol: algorithms,” IEEE J. Sel. Areas Commun., vol. 19, no. 12, pp. 2486–2494, Dec. 2001.

[20] A. Bertrand, “Applications and trends in wireless acoustic sensor networks: A signal processing perspective,” in Proc. IEEE Symp. Commun. Veh. Technol., pp. 1–6, Nov. 2011.

[21] C. Blandin, A. Ozerov, and E. Vincent, “Multi-source TDOA estimation in reverberant audio using angular spectra and clustering,” EURASIP J. Adv. Signal Process., vol. 12, no. 8, pp. 1950–1960, 2012.

[22] D. G. Knapp and C. C. Carter, “The generalized correlation method for estimation of time delay,” IEEE Trans. Acoust., Speech, Signal Process., vol. 24, no. 4, pp. 320–327, Aug. 1976.

[23] J. Chen, Y. Huang, and J. Benesty, “Time delay estimation,” in Audio Signal Processing for Next-Generation Multimedia Communication Systems. Boston: Kluwer Academic Publishers, 2004, pp. 197–227.

[24] J. J. Tashev, Sound Capture and Processing, ser. Practical Approaches. John Wiley & Sons, 2009.
13

[25] M. Cobos, F. Antonacci, L. Comanducci, and A. Sarti, “Frequency-sliding generalized cross-correlation: A sub-band time delay estimation approach,” IEEE/ACM Trans. Audio, Speech, Lang.Process., vol. 28, pp. 1270–1281, 2020.

[26] J. Chen, J. Benesty, and Y. Huang, “Robust time delay estimation exploiting redundancy among multiple microphones,” IEEE Trans. Speech Audio Process., vol. 11, no. 6, pp. 549–557, Nov. 2003.

[27] J. Benesty, J. Chen, and Y. Huang, “Time-delay estimation via linear interpolation and cross correlation,” IEEE Trans. Speech Audio Process., vol. 12, no. 5, pp. 509–519, 2004.

[28] 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.

[29] 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.

[30] F. Viola and W. F. Walker, “Computationally efficient spline-based time delay estimation,” IEEE Trans. Ultrason., Ferroelectr., Freq. Control, vol. 55, no. 9, pp. 2084–2091, Sept. 2008.

[31] B. Qin, H. Zhang, Q. Fu, and Y. Yan, “Subsample time delay estimation via improved GCC PHAT algorithm,” in Proc. Int. Conf. Signal Process., pp. 2579–2582, Oct. 2008.

[32] 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.

[33] V. Martin, K. Jelena, and G. V. K., Foundations of Signal Processing, Cambridge: Cambridge Univ. Press, 2014.

[34] H. Ogawa, “Sampling theory and Isao Someya: A historical note,” Sampling Theory in Signal and Image Process., vol. 5, no. 3, pp. 247–256, Sept. 2006.

[35] L. Wang and S. Doclo, “Correlation maximization-based sampling rate offset estimation for distributed microphone arrays,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 24, no. 3, pp. 571–582, Jan. 2016.

[36] J. C. Kiefer, “Sequential minimax search for a maximum,” Proc. American Mathematical Society, vol. 4, no. 3, pp. 502–506, Jun. 1953.

[37] K. Yamaoka, R. Scheibler, N. Ono, and J. Wakahayashi, “Sub-sample time delay estimation via auxiliary-function-based iterative updates,” in Proc. IEEE Workshop Appl. Signal Process. Audio Acoust., pp. 125–129, Oct. 2019.

[38] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, “Multichannel signal enhancement algorithms for assisted listening devices: exploiting spatial diversity using multiple microphones,” in IEEE Signal Process. Magazine, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[39] K. Lange, MM Optimization Algorithms. Philadelphia: SIAM-Society for Industrial and Applied Mathematics, 2016.

[40] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in Proc. Neural Info. Process. Syst., pp. 556–562, Jan. 2000.

[41] H. Kameoka, T. Nishimoto, and S. Sagayama, “A multipitch analyzer based on harmonic temporal structured clustering,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 15, no. 3, pp. 982–994, Mar. 2007.

[42] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” in Proc. IEEE Workshop Appl. Signal Process. Audio Acoust., pp. 189–192, Oct. 2011.

[43] A. Brendel and W. Kellermann, “Accelerating auxiliary function-based independent vector analysis,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., pp. 496–500, Jun. 2021.

[44] M. G. McGaffin and J. A. Fessler, “Algorithmic design of majorizers for large-scale inverse problems,” arXiv:1508.02938, 2015.

[45] R. Scheibler, E. Bezram, and I. Dokmanic, “Pyroomacoustics: A Python package for audio room simulations and array processing algorithms,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., pp. 351–355, Apr. 2018.