Template-Based Method for Compensation of Time Difference of Arrival in Passive Sound Source Localization under Reverberant and Noisy Environments

Yuuki Tachioka and Tomohiro Narita

Information Technology R & D Center, Mitsubishi Electric Corporation
5-1-1 Ofuna, Kamakura, Kanagawa 247-8501, Japan
E-mail: {Tachioka.Yuki@eb,Narita.Tomohiro@ct}.MitsubishiElectric.co.jp

Abstract Source localization techniques are important and effective for various applications. Without reverberation and noise, some conventional source localization techniques have achieved high accuracies. However, reverberation degrades their performance due to reflected sounds because they compare observed time differences of arrival (TDOAs) with theoretical TDOAs, which are derived from anechoic conditions and do not agree with actual ones under reverberant conditions. We propose a template-based method that compensates the discrepancy between theoretical TDOAs and actual TDOAs in order to reduce the influence of the proposed method. In this study, two types of experiments are performed to validate the effectiveness of the proposed method. The first one is a simple case that investigates the estimation accuracy in detail and the second one investigates a more practical case in a more complicated situation. For both cases, our proposed method can calibrate these errors effectively without increasing the computational time and improve the performance of conventional methods.

Keywords: sound source localization, calibration, reverberation, noise robustness, time difference of arrival (TDOA)

1. Introduction

Sound source localization techniques expand the applicability of various applications. One of the applications is surveillance [1]. Source localization based on sound is suitable for low-cost surveillance because switching the camera to the estimated sound direction widens the covering area of surveillance. Such source localization techniques are classified into passive and active ones. Passive techniques, which only use receivers and this paper deals with, are more practical but more complicated than active ones, which use both receivers and transmitters.

For the case that the target is limited to direction estimation, high accuracies have been achieved [2–4]. In addition to the direction, if the source position can be estimated, it broadens the possibility of applications using localization. Direction estimations have one unknown variable under the plane wave assumption, whereas source position estimations have two or three unknown variables under the spherical wave assumption. The latter is much more difficult than the former. The experiments reported in this paper show that the tolerance errors of the latter estimation are much smaller than those of the former estimation. In this paper, source localization is limited to the horizontal plane because vertical (height) estimation is less important.

Recently, some source localization techniques have been proposed [5,6]. Without reverberation and noise, performance is high; however, this paper shows that reverberation and noise degrade the performances of the conventional methods due to reflected sounds and measurement errors. For passive systems, to reduce their influence, some calibrations are needed [7]. Source localization methods compare an observed time difference of arrival (TDOA) with a reference TDOA. Reference TDOAs are based on the assumptions above and assume no reflected sounds and measurement errors; thus, errors degrade the source localization accuracies. To address this problem, we pro-

1One of the applications that need height localization is a robot. For height localization, microphones located at different heights are needed and this is a frequent setting for robot applications. However, a general microphone array has microphones that are located at the same height and, for these applications, height localization is less useful.
pose a template-based method that modifies reference TDOAs according to reference measurements.

Section 2 describes two conventional methods, which are the two-dimensional-cross-spectrum phase (2D-CSP) method [5] and the multichannel CSP method [6]. We proposed a template-based method in Sect. 3. Section 4 describes two source localization experiments under reverberant and noisy environments, for simple and more practical cases. The experiments show that the performances of the conventional methods are low and that our proposed method improves them.

2. Conventional Source Localization Methods

Almost all the conventional source localization methods are based on a spherical wave assumption. Although there are many source localization methods, Hayashida et al. [6] showed that the 2D-CSP method and multichannel CSP method have high accuracies. We employ these two methods as a baseline, but our proposed template-based method can be applied to any methods\(^2\).

2.1 Spherical wave assumption

In near fields where the condition
\[
\rho < \frac{2D^2}{\lambda}
\]
ist satisfied, sounds propagate as a spherical wave; otherwise, sounds propagate as a plane wave [8]. Here, \(\rho\) is the distance from the center of the microphone array to the source, \(D\) is the maximum width of the microphone array and \(\lambda\) is the wavelength. In this case, sources are assumed to be point sources, and at points with the same distance from the source, the phases are identical. For example, for a frequency of 1 kHz, \(\rho = 0.52\) [m] when \(D = 0.5\) [m] and \(\rho = 2.1\) [m] when \(D = 0.6\) [m]. When the source position is \(s\) and the position of the nth microphone among \(N\) microphones \((1 \leq n \leq N)\) is \(r_n\), the TDOA \(\tau_{n1,n2}^{\text{sph}}\) between microphones \(n_1\) and \(n_2\) is represented as
\[
\tau_{n1,n2}^{\text{sph}}(s) = \frac{d_{n1} - d_{n2}}{c} = \frac{|r_{n1} - s| - |r_{n2} - s|}{c}
\]
where \(c\) is the sound speed and \(d_n\) is the distance from the source to the \(n\)th microphone.

2.2 2D-CSP method

The CSP method estimates the TDOA \(\tau\) between two input signals from a cross spectrum of these signals [2]. First, CSP coefficients are calculated as
\[
C = \mathcal{F}^{-1} \left( X_{n1}(i) \otimes X_{n2}(i)^* \right) \\
\left( |X_{n1}(i)|^2 \right)^{-1}
\]
where \(X_n(i) = \{ X_n(i, 1), \ldots, X_n(i, k), \ldots, X_n(i, K) \}\) is the spectrum of the \(n\)th microphone input at the \(i\)th frame and \(k\)th frequency bin \((1 \leq k \leq K)\); \(\mathcal{F}\) is a short-time Fourier transform; \(*\) denotes the complex conjugate and \(\otimes\) denotes the element-wise multiplication of two vectors. Second, the TDOA \(\tau_{n1,n2}^{\text{csp}}\) is obtained as the optimal solution of the following problem:
\[
\tau_{n1,n2}^{\text{csp}} = \arg \max (\mathcal{C})
\]
The 2D-CSP method estimates \(s\) under the spherical wave assumption [5]. To do this, the number of microphones, \(N\), must be 3 or more. For example, there are two microphone pairs: \(\varphi(1) = \{1, 2\}\) and \(\varphi(2) = \{3, 4\}\). Here, for simplicity, the microphone intervals are the same. In the plane-wave case, \(|d_1 - d_2| = |d_3 - d_4|\); thus, there is no difference in the TDOA between the microphone pairs. On the other hand, in the spherical-wave case, using the difference between \(|d_1 - d_2|\) and \(|d_3 - d_4|\), the distance to the sources can be estimated. The theoretical TDOAs \(\tau_{n1,n2}^{\text{sph}}\) are determined using Eq. (2), whereas experimentally, the TDOAs \(\tau_{n1,n2}^{\text{csp}}\) can be obtained by the CSP method using Eq. (4). Some candidate source points \(S\) are prepared in advance. For each candidate point \(s \in S\), a cost function \(P(s)\) is calculated by adding the differences between the theoretical TDOAs and the observed TDOAs of \(M\) microphone pairs \((2 \leq M \leq N_C)\). If the theoretical TDOAs are near to the observed TDOAs, the cost function \(P\) will be small. If the measurement errors are sufficiently small, the source positions can be estimated at the minimum cost of \(P(s)\) as
\[
\text{arg min}_{s \in S} P(s) = \text{arg min}_{s \in S} \sum_{n=1}^{M} \left( \tau_{n1,n2}^{\text{sph}}(s) - \tau_{n1,n2}^{\text{csp}}(s) \right)^2
\]
where \(\varphi(m)\) is the \(m\)th microphone pair. Note that because one microphone pair can only indicate that a sound source exists on a hyperbola, two or more different microphone pairs (i.e., three or more microphones) are needed.

2.3 Multichannel CSP (M-CSP) method

The 2D-CSP method adds the differences in the TDOAs for each microphone pair, whereas the M-CSP method considers the differences in the TDOAs of all microphone pairs simultaneously by calculating the all-pair correlation matrix as
\[
R_k = \begin{bmatrix}
\xi_{1,1,k} & \cdots & \xi_{1,N,k} \\
\vdots & \ddots & \vdots \\
\xi_{N,1,k} & \cdots & \xi_{N,N,k}
\end{bmatrix}
\]
and compares this correlation matrix with given steering vectors [6]. This simultaneous consideration of the
correlation between each microphone pair improves the accuracy. Each component is represented as

\[
[H_{1,n}, H_{2,n}, \ldots, H_{N,n}^*] = \frac{X_{n,i}^* X_{n,i}^*}{|X_{n,i}||X_{n,i}^*|}
\]

where \( \tau \) denotes the transpose.

The steering vector \( a_k(s) \) for \( s \) is obtained as

\[
a_k(s) = \left[ e^{-j\omega_k|\tau_1 - \tau_2|/c}, \ldots, e^{-j\omega_k|\tau_{N-1} - \tau_N|/c} \right]^T
\]

where \( j \) is the imaginary unit and \( \omega_k \) is the \( k \)th angular frequency. For each \( s \),

\[
P_k(s) = \frac{1}{a_k(s)^H R_k a_k(s)}
\]

is calculated where \( R_k \) is the Hermitian transpose.

If \( s \) is near to the actual source position, \( P_k(s) \) becomes small. After averaging \( P_k(s) \) over the target frequency bins (\( k_L \leq k \leq k_H \)), the source positions can be estimated as

\[
\arg\min_{s \in S} P(s) = \arg\min_{s \in S} \frac{k_H - k_L}{\sum_{k=k_L}^{k_H} P_k(s)}
\]

The M-CSP method outperforms the 2D-CSP and 2D-multiple signal classification (MUSIC) methods [6].

### 3. Template-Based Method

In real situations with reverberation, the theoretical TDOA and the observed TDOA for the correct positions can differ due to reverberation or measurement errors. To reduce their influence, we propose a template-based method whose cost function \( P \) is given in a generalized form as

\[
\arg\min_{s \in S} P(s) = \arg\min_{s \in S} \sum_{m=1}^{M} \left( \tau_{r \Phi(s)} - \tau_{o \Phi(s)} \right)^2
\]

where \( \tau_{r \Phi(s)} \) is the reference TDOA for position \( s \) and \( \tau_{o \Phi(s)} \) is the observed TDOA. The 2D-CSP method uses \( \tau_{sph} \) for reference and \( \tau_{csp} \) for observation. Source localization methods use the theoretical TDOA as a reference, but observations generally contain errors. For example, reflected waves have high correlations with direct waves, which leads to TDOA estimation errors. Figure 1 shows the errors caused by reverberation. In this case, the observed TDOA \( \tau_{csp} \) is longer than the theoretical one \( \tau_{r \Phi(s)} \) when the direct wave for the first microphone and the reflected wave for the second microphone have higher correlations than direct waves. The errors between the theoretical TDOA and observed TDOA are denoted as \( \epsilon \). The reference TDOA is modified by the errors \( \epsilon \), which are calculated for known positions \( s \in S \) in the reference measurements a priori after \( \tau_{o \Phi(s)} \) is calculated as Eq. (12).

\[
\epsilon_m(s) \leftarrow \tau_{o \Phi(s)} - \tau_{sph}(s)
\]

In the 2D-CSP case, this formula is

\[
\epsilon_m(s) \leftarrow \tau_{csp}(s) - \tau_{sph}(s)
\]

Prior to the first use, these errors \( \epsilon \) are calculated for the target points \( s \) and stored as a template. These modified references are expected to cancel out the errors. In the case without errors or reflections, \( \epsilon \) is zero and the proposed method exactly matches the original method.

By considering the stored \( \epsilon \) for position \( s \), we use the reference below, \( \tau_{r \Phi(s)} \), instead of \( \tau_{sph} \):

\[
\tau_{r \Phi(s)} \approx \tau_{sph}(s) + \epsilon_m(s)
\]

This method can be applied to any source localization method, no only the 2D-CSP method.

### 4. Source Localization Experiments in Simple Case

#### 4.1 Setup

To validate the effectiveness of the proposed method, simple experiments were performed. The number of sound sources was 25 in the room shown in Fig.2 and 25 corresponding impulse responses were recorded. The notation \( D\{\bar{D}\} \{R\} \) denotes that the source exists in direction \( D \) [°] and at a distance \( R \) [cm] from the center of the microphone array, which is the origin. To construct evaluation data, the impulse responses were convolved with clean speech utterances, which were composed of a few words each. Vocabularies were control words for air conditioners, such as “Set to 30 degrees.”. The room reverberation time measured at the room center was 580 ms, where there were rich reflections in this room. We validated two cases: a reverberant condition in Sect. 4.2 and a reverberant and noisy condition, where the air conditioner noise was added at a signal-to-noise ratio (SNR) of 12 dB,
in Sect. 4.3. The sampling frequency was 16 kHz\(^3\), and the window length and frame shift of the short-time Fourier transform were 60 ms and 30 ms, respectively. Frequency bands higher than 150 Hz were used. The candidate points of the sound source were 25 positions, which were the same as the above source positions. Seven microphones were prepared for the recordings as shown in Fig. 2. In this case, the sound speed \(c\) and the microphone positions \(r_i\) are constant and known. We show the results using three microphones (chs. 1, 2 and 5) (abbreviated to pair-3ch) and five microphones (chs. 1, 3, 5, 6 and 7) (abbreviated to pair-5ch). The microphone pairs were all the pairs for all the methods \((M = N C_2)\). For our proposed method, the reference measurements were 10 utterances at each point by one female speaker, who was different from the evaluation speakers.

The estimation performances were evaluated on the basis of the distance between the actual source position \(s_a\) and the estimated source positions \(s_e\) in terms of two measures: the estimation accuracy within 25% tolerance (shown in a bar graph) [%] (i.e., the ratio of the number of cases where \(|s_a - s_e|/|s_a|\) is less than 0.25 to the total number of cases) and the average absolute error (shown in a line graph) [m] \(|s_a - s_e|\).

Two types of evaluation are necessary because, for the former one, farther sources have larger tolerance errors, whereas for the latter one, nearer sources have larger tolerance errors relative to the actual distance. Tolerance errors are different for different applications but 30 cm errors can be tolerated for many applications. One of the examples is home appliances such as air conditioners that detect humans and concentrate the air flow to that area. In this kind of application, 30 cm errors are acceptable.

4.2 Results and discussion (reverberant condition)

First, we compared two conventional methods: the 2D-CSP and M-CSP methods. Figure 3 shows the av-
The value of the cost function $P$ in Eq. (5) is shown in Fig. 6. The source is located at “D060R300”.

Fig. 6 Value of cost function $P$ in Eq. (5): The source is located at “D060R300”.

The contours of estimation accuracy [%] of the 2D-CSP+template method are shown in Fig. 7. The estimation accuracies in pair-5ch were higher than those in pair-3ch under reverberant conditions.

Fig. 7 Contours of estimation accuracy [%] of the 2D-CSP+template method with ±25% tolerance using pair-5ch under reverberant condition

The average estimation accuracy and error for the 2D-CSP method are shown in Fig. 8. The computational time was normalized by that of the pair-3ch case for the 2D-CSP method.

Fig. 8 Computational time normalized by that of the pair-3ch case for the 2D-CSP method

The performance of the M-CSP method was higher than that of the 2D-CSP method as shown in [6]. Figure 4 shows contours of the estimation accuracies of the 2D-CSP method, which show that there were many points that could not be estimated. Compared with the 2D-CSP method, the M-CSP method improved the estimation accuracies on average; however, Fig. 5 shows that there were still many points that could not be estimated.

To evaluate the tolerance errors of direction estimation and source localization, Fig. 6 shows the value of the cost function $P$ (pair-5ch). The differences in $P$ between different distances in the same direction were much smaller than those between directions for the same distance. This shows the difficulty of 2D-source localization.

Figure 3 also shows the result of our method. We validated the effectiveness of our ‘2D-CSP+template’ method. Figure 7 shows that in pair-3ch, there were some points that had low accuracies, but, in pair-5ch, almost all the points had high accuracies of over 90%. Calibrations effectively reduced the influence of reflected sounds and measurement errors.

Figure 8 shows a comparison of the computational times of the above methods. All the methods were implemented by C++ and computational times were obtained using the same computer. Computational times were normalized by that of the 2D-CSP method. The computational time of the M-CSP method was 30 times longer than those of the 2D-CSP and 2D-CSP  + template methods.
Table 1  Localization and speech detection results for the development set and test set of the DIRHA corpus: The performance criteria for source localization are the fine error (FE), gross error (GE) and percentage of correct localization (PCor). The unit of errors is mm.

<table>
<thead>
<tr>
<th>Method</th>
<th>Development set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FE</td>
<td>GE</td>
</tr>
<tr>
<td>2D-CSP</td>
<td>298</td>
<td>602</td>
</tr>
<tr>
<td>2D-CSP+template</td>
<td>303</td>
<td>592</td>
</tr>
<tr>
<td>M-CSP</td>
<td>347</td>
<td>1307</td>
</tr>
<tr>
<td>SRP-PHAT</td>
<td>289</td>
<td>826</td>
</tr>
</tbody>
</table>

4 CSP+template methods. Those of the pair-5ch case were two or three times larger than those of the pair-3ch case, which were approximately proportional to the number of pairs from 3 to 10.

4.3 Results and discussion (reverberant and noisy condition)

Figure 9 shows the accuracies under a reverberant and noisy environment. By comparison with Fig.3, it can be seen that noise degraded the estimation performance. The 2D-CSP method had unsatisfactory performance but the 2D-CSP+template method improved the performance. The proposed method improved the performance less than that for the reverberant case because the template was built for the reverberant case (i.e., without noise). Figure 10 shows the contours of the estimation accuracies of the 2D-CSP+template method. For almost all the points, the accuracy was more than 80%. Even without noise, the conventional method was not accurate at some points as shown in Fig.4. The proposed method is much more practical than the conventional method. We thus validated the effectiveness of our 2D-CSP+template method for a reverberant and noisy condition.

5. Source Localization Experiments in Practical Case

5.1 Setup

In addition to the experiments in Sect. 4, a more practical and complicated case was examined. Figure 11 shows the setup of the experiment. To simulate voice-active home appliances, synchronously recorded sound files (approximately 1-2 min) were provided by the DIRHA consortium [9]. To simulate realistic environments, these databases were recorded in a real house, which consisted of five rooms: a kitchen, living room, corridor, bathroom and bedroom. Localization was limited to the kitchen and living room and, for these rooms, a circular six-microphone array was installed at the center of the room. Additionally, for all rooms, several two- or three-microphone arrays were installed on the walls encompassing the room. In total, 40 microphones were used. Microphone pairs were selected within each array because microphones belonging to separate arrays were far apart and their correlations were too small.

A development set (dev) and a test set (test) were provided. According to the regulations, any parameters in the dev set can be tuned. Both sets consist of Real and Simulations subsets. In the Real set, for each task, there is only one speaker in one room, who moves around the room. To simulate the dialog between the speaker and system, the replies of the system sometimes break in, but they are provided separately. In the Simulations set, there can be multiple speakers in different rooms but they are still. The system performance was evaluated using the provided evaluation tools.

We focused on 2-D localization because height localization is less important than horizontal localization as mentioned in the introduction. The speech data were downsampled from the original 48 kHz to 16 kHz for our experiments. The frame size was

\[\text{frame size} = 4\text{ frames} = 256\text{ samples}\]

\[\text{rate} = 16\text{ kHz}\]

\[\text{duration} = 100\text{ ms}\]
960 and the frame shift was 800. We compared the performances of the 2D-CSP and 2D-CSP+template methods with those of the M-CSP [6] and the SRP-PHAT\textsuperscript{5,6} [10] methods. Fine errors were defined as localization errors of less than 50 cm. These tasks assume that the source position and voice activity area need to be simultaneously estimated. However, for focusing on the comparison of sound localization, in this case, the correct speech area was given.

5.2 Results and discussion

Table 1 shows the results. The performance of the 2D-CSP method was higher than those of the M-CSP and SRP-PHAT methods. Moreover, the computational complexity was much lower than those of the M-CSP and SRP-PHAT methods. The 2D-CSP method is the most practical among these methods. The proposed 2D-CSP+template method improved the performance of the 2D-CSP method significantly, proving its effectiveness for localization in reverberant environments. In this practical and more complicated situation, the proposed method can improve the estimation accuracy.

6. Conclusion

We proposed a template-based method in order to reduce the influence of reflected sounds and measurement errors, which degrade the source localization performance. Without increasing the computational time, our method can improve source localization accuracies for reverberant and noisy environments.

References


\textsuperscript{5}http://www.lems.brown.edu/array/tools/srplems.m
\textsuperscript{6}A long frame size (1 second) was used.


Yuuki Tachioka is a researcher at Information Technology R&D Center, Mitsubishi Electric Corporation. In 2008, he joined the R&D Center after he graduated from the School of Engineering, the University of Tokyo, in 2006 and the Graduate School of Frontier Sciences of the same university in 2008. His research interests include speech recognition, speech enhancement and array signal processing. He is a member of the Acoustical Society Japan (ASJ), the Mathematical Linguistic Society of Japan and the Information Processing Society of Japan. He received the Excellent Master’s Thesis Award from the Architectural Institute of Japan in 2008 and the Awaya Award from ASJ in 2014.

Tomohiro Narita is a group manager of the language processing group at Information Technology R&D Center, Mitsubishi Electric Corporation. In 1997, he joined the R&D Center after he graduated from the School of Engineering, Tokyo Institute of Technology, in 1995 and the Graduate School of the same university in 1997. He has been engaged in the research and development of speech recognition.

(Received June 21, 2016; revised November 28, 2016)