Localization of Multiple Sound Sources Based on Inter-Channel Correlation Using a Distributed Microphone System

Kook Cho, Hajime Okumura, Takanobu Nishiura and Yoichi Yamashita

College of Information Science and Engineering, Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu, Shiga, 525-8577 Japan
{cho@slp.is, okumura@slp.is, nishiura@is, yama@media}.ritsumei.ac.jp

Abstract
Recently the importance of hands-free speech interfaces is increasingly recognized. However, in real environments, the presence of ambient noises and room reverberations seriously degrades the performance of the hands-free speech recognition. Reliable sound source localization is necessary to maximize the effect of noise reduction. This paper proposes a new method of multiple sound source localization using a distributed microphone system that is a recording system with multiple microphones dispersed to a wide space. The proposed method localizes a sound source by finding the position that maximizes the accumulated correlation coefficient between multiple channel pairs. After the estimation of the first sound source, a model of the accumulated correlation for a single sound source is subtracted from the observed distribution of the accumulated correlation. Subsequently, the second sound source is searched again. To evaluate the effectiveness of the proposed method, experiments of multiple sound source localization were carried out in an actual office room. The result shows that multiple sound source localization accuracy is about 96.0%. The proposed method could realize the multiple sound source localization robustly and stably.

Index Terms: sound source localization, distributed microphone system, TDOA, real environment, multiple sound sources

1. Introduction
The high-quality sound capture is very important for hands-free speech interfaces, because ambient noises and room reverberations seriously degrade the quality of captured speech signals in real noisy environments. To suppress noises and reverberations and extract the sound of sources, sound source localization is one of key techniques.

A microphone array system enables estimation of the direction of arrival (DOA) of the observed speech signal based on the time delay of arrival (TDOA) between multiple captured signals. The Minimum Variance (MV) method [1], the Multiple Signal Classification (MUSIC) method [2], the Cross-Correlation (CC) method [3] and the Cross-power Spectrum Phase (CSP) analysis [3][4][5][6] are popular DOA estimation methods. The MV beamformer method identifies the speech signal in the direction where the spectrum entropy of the signal is minimized. The MUSIC method is a kind of DOA estimation algorithm based on eigenvector decomposition, which is based on a subspace-based method. The MV and the MUSIC method can estimate only M – 1 DOAs with an M-element microphone array.

A sound source can be theoretically localized using two sets of microphone array by combining two independent directions. However, this approach degrades the performance of sound source localization in noisy and reverberant environments, because small errors of direction estimation may result in a large error of position estimation. In addition, the MV and the MUSIC method are very difficult to process in real time because of its heavy computational load and complexity. Although beamforming can give good results, the computation is generally too expensive to allow the likelihoods to be computed at all possible locations. Although time-delay estimation methods are fast, they generally perform poorly in highly reverberant environments. Recently an accumulated correlation algorithm [7][8][9][10][11] was proposed to combine the advantages of these two approaches. Instead of taking the peak of each correlation vector, all the correlation values from all the vectors are accumulated in a common coordinate system. A distributed microphone system that disperses multiple microphones in a wide space provides a possibility of sound source localization with a high accuracy and small computational costs. The distributed microphone system can directly localize sound sources with optimization for a wide range of the space.

More than one sound source may exist in real environments. Several persons sometimes talk simultaneously in meeting or discussion. A technique of the sound source localization is required to work in multiple sound source situations. To solve this problem, this paper proposes a new method of multiple sound source localization using a distributed microphone system that is widely distributed and placed under the ceiling of a room. A number of pairs of microphones give a series of correlation coefficients as a function of the time delay. The sound source is localized based on correlation of two channel signals that are delayed with the time delay of arrival for a hypothetical sound source. The correlation coefficients for hypothetical sound sources are accumulated over many microphone pairs. The proposed method localizes a sound source by finding the position that maximizes the accumulated correlation coefficient between multiple channel pairs. After the estimation of the first sound source, the accumulated correlation for a single sound source is subtracted from the observed distribution of the accumulated correlation. Multiple sound sources positions can be estimated by adapting the proposed method repeatedly.

2. Sound source localization method
2.1. Estimation of TDOA with the CSP method
The Cross-Correlation (CC) method [3] localizes a sound source by utilizing CC coefficients based on cross-power spectrum between captured signals. However, it is not enough robust, because cross-power spectrum is directly affected by the amplitude of noise signals. To overcome this problem, Cross-power Spectrum Phase (CSP) analysis [3][4][5][6] has been proposed as an advanced CC method. It can accurately
localize the sound sources without dependence on spectral characteristics of captured signals, because it only utilizes phase difference between captured signals with a pair of microphones by employing normalized cross-power spectrum with the amplitude of captured signals. The CSP method is widely used because of their computational efficiency and stability. The direction of the sound source can be obtained by estimating a TDOA between two microphone outputs. The CSP coefficients are calculated by the following equation.

$$CSP_{ij}(k) = \text{IDFT} \left[ \frac{\text{DFT}[s_i(n)] \text{DFT}[s_j(n)]^*}{|\text{DFT}[s_i(n)]| |\text{DFT}[s_j(n)]|} \right]$$  \hspace{1cm} (1)$$

$$\tau = \arg\max_k (CSP_{ij}(k))$$ \hspace{1cm} (2)

where $s_i(n)$ and $s_j(n)$ are the signals acquired through the $i$-th and $j$-th microphones, $n$ and $k$ are the time index, $\text{DFT}[\cdot]$ is the discrete Fourier transform, $\text{IDFT}[\cdot]$ is the inverse discrete Fourier transform, the symbol * is the complex conjugate, $CSP_{ij}(k)$ is the CSP coefficients, and $\tau$ is an estimated TDOA. The TDOA can be estimated by finding the maximum value of the CSP coefficients.

### 2.2. Sound sources localization based on accumulated the inter-channel correlation

In real environments, the presence of ambient noises and room reverberations seriously degrades the accuracy in sound sources localization. This paper proposes a new localization algorithm for multiple sound sources based on the inter-channel correlation calculated by the CSP method.

#### 2.2.1. Single sound source localization

The procedure for single sound source localization by the inter-channel correlation method is as follows:

1. Make a set of hypothetical sound sources.
2. Calculate each TDOA using a transmission path between a hypothetical sound source and two microphones.
3. The correlation coefficients between two channel signals delayed with the TDOA are accumulated over all the microphone pairs.
4. The sound source is localized as the hypothetic position that maximizes the accumulated correlation coefficients. The correlation coefficients for many microphone pairs are calculated by the CSP method. A TDOA, $k_{ijp}$, between the $i$-th and $j$-th microphones for the $p$-th hypothetical sound source is derived from the equation (3).

$$k_{ijp} = \frac{|m_i - s_p| - |m_j - s_p|}{c}$$ \hspace{1cm} (3)

where $m_i$ is the position coordinate of the $i$-th microphone, $s_p$, $p=1, 2, ... , P$) is the $p$-th hypothetical sound source position coordinate, $c$ is the sound propagation speed. Then the accumulated CSP coefficient in the $p$-th hypothetical sound source, $CSP_{acc}(p)$, is derived from the equations (4) and (5).

$$CSP_{acc}(p) = \sum_{i,j \in S} CSP'_{ij}(k_{ijp}, w)$$ \hspace{1cm} (4)

$$CSP'_{ij}(k_{ijp}, w) = \max [CSP_{ij}(k)]$$ \hspace{1cm} (5)

where $CSP_{ij}(k)$ is the CSP coefficient of the $i$-th and $j$-th microphone pair for TDOA, $k$, as shown in the equation (1). $CSP'_{ij}(k_{ijp}, w)$ is an adjusted correlation coefficient that is defined as a maximum CSP coefficient among the delay range near to $k_{ijp}$. The delay $k_{ijp}$ is a theoretical value of the time delay between the $i$-th and $j$-th microphone pair for the $p$-th hypothetical sound source, and it is calculated based on the microphone positions, shown as the equation (3). $CSP'_{ij}(k_{ijp}, w)$ is accumulated in steady of a raw CSP correlation, $CSP_{ij}(k_{ijp})$, to consider measurement errors of the microphone positions. The parameter, $w$, is a CSP window that is a search range in time domain to find a maximum CSP coefficient. The CSP window is defined as $[k_{ijp}-w, k_{ijp}+w]$. $S$ is a set of microphone pairs. The sound source positions can be estimated by finding the maximum values of the accumulated CSP coefficients by the equation (6).

$$\hat{l} = \arg\max_p (CSP_{acc}(p))$$ \hspace{1cm} (6)

where $\hat{l}$ is the estimated position of sound source. This procedure is schematically shown in Figure 1.

#### 2.2.2. Multiple sound sources localization

The multiple sound source positions can be repeatedly estimated by finding the maximum values of the accumulated correlation coefficients in descending order. But, the accumulated correlation peak for the second sound source is not necessarily the second largest peak in observed accumulation correlation distribution, $CSP_{acc}(p)$, shown in Figure 2 (a). Figure 2 illustrates examples of accumulated correlation distribution in a one-dimensional space. In Figure 2 (a), the second largest peak of the accumulated correlation does not locate in the second sound source, $l_2$, but also in the neighbor of the first sound source, $l_1$. The proposed method introduces the subtraction of the accumulated correlation in order to avoid such a localization error of the second sound source. The average distribution of the accumulated correlation is obtained with training data of accumulated
correlation distribution for a single sound source, and it is called as the Single Source model (SS-model) which is denoted by \( \text{Peak}(p) \). The SS-model is normalized so that the peak of correlation is 1. The accumulated correlation distribution for multiple sound sources is modified by the subtraction of the SS-model. The modified distribution, \( \text{CSP}'_{\text{acc}}(p) \), is calculated by

\[
\text{CSP}'_{\text{acc}}(p) = \text{CSP}_{\text{acc}}(p) - \text{CSP}_{\text{acc}}(\hat{l}_1)\text{Peak}(p - \hat{l}_1) \quad (7)
\]

using the estimated position of the first sound source, \( \hat{l}_1 \), and the SS-model. Figure 2 (c) shows an example of the modified distribution of the accumulated correlation. The second source can be successfully identified by finding a correlation peak in the modified distribution since the peak of the first sound source was removed by the subtraction of the SS-model. The estimated position of the second sound source, \( \hat{l}_2 \), is obtained by

\[
\hat{l}_2 = \arg\max_{p \in P} \{ \text{CSP}'_{\text{acc}}(p) \} \quad (8)
\]

In the case of more than two sound sources, sound source positions can be repeatedly estimated by modifying the accumulated correlation distribution based on the earlier estimated sound position and the SS-model subtraction.

### 3. Experimental evaluation

#### 3.1. Experimental conditions

We recorded data in an actual room and evaluated the effectiveness of the proposed method. Figure 3 shows the layout of sound sources and microphones in an experimental environment. The number of the microphones is 16 in our distributed microphone system which is installed in a 4x4 lattice condition under the ceiling. The distance between the microphones was 135 cm, and the height of the microphones was 233 cm. As shown in Figure 3, several noise sources such as a server and workstations existed in the experimental environment. Room reverberation (\( T_{60} \)) was 0.39 sec and ambient noise level was 48.2 to 56.0 dBA. Thus, this room is a higher noisy environment.

The number of the sound sources is 2 in this experiment. Speech materials consist of four Japanese sentences spoken by a male speaker and a female speaker, and they are played through two loudspeakers and recorded by the distributed microphone system. Direction of the loudspeakers was set to one of four directions: north, east, south, and west. Angle of the loudspeakers was set to the horizontal direction. Eight positions indicated by small black circles in Figure 3 are evaluated as speech source positions. Two loudspeakers are
put in two positions which are selected among the eight speech source positions. Thus, changing the setting of the loudspeaker, we recorded the data of 96 sentences in total. The sampling frequency was 16[kHz], and quantization was 16[bits]. We tried to evaluate the proposed method with 1,024[msec] frame length. Sound source localization estimation is conducted for speech periods with 100[msec] shift interval. The height of the loudspeaker was 108[cm]. A speech source is localized under a condition that the height of the source is given. We investigated the accuracy of sound source localization for 2-dimensional lattices of hypothetical speech sources; 7.5cm (3481 point), as shown in Figure 4.

3.2. Experimental results

The performance of the proposed method is controlled by the CSP window, \( w \). Figure 5 shows the correct estimation rate of multiple sound sources. The correct estimation is defined that the distance between the estimated and the correct positions is less than 20[cm]. Figure 6 shows the average error distances between the correct and the estimated sound sources. In Figure 5 and 6, the method 1 indicates results of the proposed method and the method 0 indicates results of the baseline method which identifies the second largest peak in the original accumulated correlation distribution as the second sound source. In Figure 5, when the CSP window, \( w \), is increased, localization accuracy is improved. However, too large \( w \) decreases the accuracy. The accuracy in the multiple sound source localization was maximized by 125[micro sec]. The result shows that multiple sound source localization accuracy is 95.7% for optimized parameters. In Figure 6, average error distances are less than 17.8cm. Even if a sound source was incorrectly localized, estimated positions were in the vicinity of the correct position. The method 1 clearly has better results than the method 0. It improves accuracy to subtract the high accumulated correlation range around a sound source from accumulated correlation of hypothetical sound source. These results show that the proposed method precisely estimates the multiple sound sources. The proposed method could estimate the multiple sound source localization robustly and stably.

4. Conclusions

This paper proposes a new multiple sound source localization method based on the accumulated inter-channel correlation using a distributed microphone system. The experiments were carried out to evaluate the proposed method in a real environment. As a result of evaluation experiments, we confirmed that the multiple sound source localization estimation performance of the proposed method is superior. In future works, introduces the subtraction of the accumulated correlation in order to reform the localization error by the reflection sound. The accumulated correlation is reformed by the subtraction of the accumulated correlation coefficient model corresponding to the reflection sound from the observed distribution. Our final goal is to acquire source separation by using results of sound source localization.

5. Acknowledgements

We would like to thank Mr. Masato Nakayama and Mr. Yuki Denda for their assistance during this work.

6. References