Multiple Sound Source Localization Based on Inter-Channel Correlation Using a Distributed Microphone System in a Real Environment

Kook CHO†, Hajime OKUMURA††, Nonmembers, Takanobu NISHIURA†, and Yoichi YAMASHITA†, Members

SUMMARY In real environments, the presence of ambient noise and room reverberations seriously degrades the accuracy in sound source localization. In addition, conventional sound source localization methods cannot localize multiple sound sources accurately in real noisy environments. 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 area. 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 typical pattern 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 two sound source localization were carried out in an office room. The result shows that sound source localization accuracy is about 99.7%. The proposed method could realize the multiple sound source localization robustly and stably.

key words: real environment, multiple sound sources, sound source localization, distributed microphone system, inter-channel correlation, TDOA

1. Introduction

Accurate localization of multiple sound sources is one of the most important issues for various speech applications. Reliable sound source localization is necessary to maximize the effect of noise reduction for hands-free speech recognition systems, online conferencing systems, security systems, and so on.

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]–[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 their heavy computational load and complexity. Although beamforming such as the MV method 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 such as the CC method and the CSP method are fast, they generally perform poorly in highly reverberant environments. Recently an accumulated correlation algorithm [7]–[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.

More than one sound source may exist in real environments. For example, several persons sometimes talk simultaneously in a meeting or discussion. A technique of the sound source localization is required to work in multiple sound source situations. To solve this problem, 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 typical pattern of the accumulated correlation for a single sound source is subtracted from the observed distribution of the accumulated correlation, and the second sound source is localized by finding the maximum correlation. The position of multiple sound sources can be estimated by adapting the proposed method repeatedly.

The remainder of this paper is organized as follows. Section 2 describes the conventional sound source localization method and the proposed method. Section 3 describes the sound source localization estimations that were carried out in an office room. Section 4 briefly describes the conclusions we reached and future works.

Copyright © 2010 The Institute of Electronics, Information and Communication Engineers
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 the CC coefficients based on cross-power spectrum between captured signals. However, it is not robust enough, because cross-power spectrum is directly affected by the amplitude of noise signals. To overcome this problem, the cross-power spectrum phase (CSP) analysis [3]–[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 its 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.

\[ CS_{ij}(k) = \text{IDFT}\left[ \frac{\text{DFT}[s_i(n)] \cdot \text{DFT}[s_j(n)]^*}{||\text{DFT}[s_i(n)]|| \cdot ||\text{DFT}[s_j(n)]||} \right], \]

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, DFT[·] is the discrete Fourier transform, IDFT[·] is the inverse discrete Fourier transform, the symbol * is the complex conjugate, \( CS_{ij}(k) \) is the CSP coefficients, and \( \tau \) is an estimated TDOA. The TDOA can be estimated by finding the maximum value of the CSP coefficients.

\[ \tau = \arg\max_k (CS_{ij}(k)), \]

Fig. 1 Concept of sound source localization with a pair of microphones.

\[ x_2(t) = \xi_2 s(t - \varphi_2) + n_2(t), \]  
\[ \tau = \varphi_2 - \varphi_1 = \frac{|d_1 - d_2|}{c}, \]

where \( \xi \), \( \varphi \), and \( n \) are an attenuation coefficient, a arrival time of \( x_1(t) \), and a noise on \( t \)-th frame, respectively. \( c \) is the sound propagation speed. Therefore, each TDOA measurement between two microphones determines that the position of sound source must lie on a hyperbola with a constant distances difference between the two microphones. For any point on a hyperbola as shown at Fig. 1 (a), the difference between distances to microphones is constant. Figure 1 (b) shows that the difference of distance from a sound source to microphones causes a TDOA. The captured signals \( x_1(t) \), \( x_2(t) \), and the TDOA \( \tau \) are related as follows:

\[ x_1(t) = \xi_1 s(t - \varphi_1) + n_1(t), \]
fixed microphones to provide an intersection point that locates the position of sound source. Intersection gives the sound source location. The sound source can be localized as a crossing point of sound directions estimated using two pairs of microphones as shown in Fig. 1 (c).

2.3 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 source localization. This paper proposes a new localization algorithm for multiple sound sources based on the inter-channel correlation calculated by the CSP method.

2.3.1 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 correlation coefficients for various TDOA from received signals in each microphone pair.
3. Calculate a TDOA using a transmission path between a hypothetical sound source and two microphones, and 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 Eq. (6).

\[
k_{ijp} = \frac{|m_i - s_p| - |m_j - s_p|}{c},
\]

where \( m_i \) is the position coordinate of the \( i \)-th microphone, \( s_p(=1, 2, \ldots, P) \) is the \( p \)-th hypothetical sound position coordinate, \( c \) is the sound propagation speed. Then the accumulated CSP coefficient in the \( p \)-th hypothetical sound source, \( CSP_{awc}(p) \), is derived from Eq. (7) and Eqs. (8), (9), (10).

\[
CSP_{awc}(p) = \sum_{(i,j) \in S} CSP'_{ij}(k_{ijp}),
\]

\[
CSP'_{ij}(k) = \max\{CSP_{ij}(k-w), \ldots, CSP_{ij}(k+(w-1)), CSP_{ij}(k+w)\},
\]

\[
CSP'_{ij}(k) = \frac{\sum_{r=-w}^{w} CSP_{ij}(k+r)}{2w + 1},
\]

\[
CSP'_{ij}(k) = \frac{w - |r|}{w} CSP_{ij}(k + r)
\]

\[
CSP'_{ij}(k) = \frac{\sum_{r=-w}^{w} CSP_{ij}(k+r)}{2w + 1},
\]

Fig. 2 Adjusted CSP correlations.

where \( CSP_{ij}(k) \) is the CSP coefficient of the \( i \)-th and \( j \)-th microphone pair for TDOA, \( k \), as shown in Eq. (1). \( S \) is a set of microphone pairs. 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 Eq. (6). \( CSP_{ij}(k_{ijp}) \) is an adjusted CSP correlation and it is accumulated instead of a raw CSP correlation, \( CSP_{ij}(k_{ijp}) \), to consider measurement errors of the microphone positions. We investigate three types of the adjusted CSP correlations which are defined by Eqs. (8), (9), and (10). Figure 2 shows these adjusted CSP correlations schematically. Eqs. (8), (9), and (10) are called the maxCSP, avgCSP, and WavgCSP, shown in Fig. 2 (a), (b), and (c), respectively. The maxCSP and avgCSP use the maximum and the average value of CSP coefficients, respectively, in the time domain search range. In case of WavgCSP, the average value of tri-linearly weighted CSP coefficients is employed as in Fig. 2 (c). The parameter, \( w \), is a CSP window width that controls a search range in time domain. The CSP window is defined as \(|k_{ijp} - t| \leq w \). The sound source positions can be estimated by finding the maximum values of the accumulated CSP coefficients by Eq. (11).
(a) The difference of distance from a hypothetical sound source to microphones causes a TDOA.

(b) Calculate correlation coefficients from received signals.

(c) The sound source is localized from an accumulated correlation coefficient.

\[ \hat{l} = \arg \max_p (CS P_{acc}(p)), \]  

where \( \hat{l} \) is the estimated position of the sound source. This procedure is schematically shown in Fig. 3. In Fig. 3 (a), p1, p2, and p3 are the hypothetical sound sources, and p2 is assumed as the correct sound source. The difference of distance from a hypothetical sound source to microphones causes a TDOA as shown in Fig. 3 (a). Figure 3 (b) shows the time-varying CSP coefficients for each microphone pair. The CSP coefficients are calculated for various TDOAs from a received signal, and they are variable according to the hypothetical positions as shown in Fig. 3 (b). The CSP coefficient in the correct position is higher than those in other hypothetical ones. Figure 3 (c) shows accumulated CSP coefficients according to different hypothetical sound sources’ positions. The sound source is localized as the hypothetic position that maximizes the accumulated correlation coefficients.

2.3.2 Multiple Sound Source Localization

The procedure for multiple sound source localization is as follows:

1. After the estimation of the first sound source, a typical pattern of the accumulated correlation for a single sound source is subtracted from the observed distribution of the accumulated correlation.

2. The second sound source is localized by finding the maximum correlation again.

3. The position of multiple sound sources can be estimated by applying the proposed method repeatedly.

Figure 4 illustrates examples of accumulated correlation distribution. In Fig. 4 (a), (b), and (c), the horizontal axes show positions in the space of hypothetical sound sources and the vertical axes show the accumulated correlation. Although the actual space of hypothetical sound sources is two-dimensional, it is drawn as a one-dimensional line just for schematic explanation in these figures. In addition, a dashed horizontal line located at the value of 1.0 on the vertical axis indicates that these coefficients were normalized so that the peak of correlation is 1. The accumulated correlation peak for the second sound source is not necessarily the second largest peak in observed accumulation correlation distribution, \( CS P_{acc}(p) \), shown in Fig. 4 (a). The second largest peak of the accumulated correlation does not locate in the second sound source, \( \hat{l}_2 \), but also in the neighbor of the first sound source, \( \hat{l}_1 \). 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 accumulated correlation distribution for a single sound source, and it is called the Single Source model (SS-model), shown in Fig. 4 (b). Figure 4 (c) shows an example of the modified distribution of the accumulated correlation.

Figure 5 illustrates examples of the SS-model of the accumulated correlation that is obtained with training data. In these figures, the horizontal and vertical axes have the same meanings as those in Fig. 4. The SS-model is obtained with the various accumulated correlation distributions for a single sound source shown in Fig. 5 (a). The SS-model is
calculated by averaging the various accumulated correlation distributions after they are normalized so that the peak of correlation is 1, as shown in Fig. 5 (b). The accumulated correlation distribution for multiple sound sources is modified by the subtraction of the SS-model shown in Fig. 5 (c). The modified distribution, \( CS \, P_{\text{acc}}'(p) \), is calculated by

\[
CS \, P_{\text{acc}}'(p) = CS \, P_{\text{acc}}(p) - CS \, P_{\text{acc}}(\hat{l}_1) \text{Peak}(p - \hat{l}_1),
\]

using the estimated position of the first sound source, \( \hat{l}_1 \), and the SS-model. \( \text{Peak}(p) \) is the correlation distribution of the SS-model. 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

\[
\hat{l}_2 = \text{argmax}_p (CS \, P_{\text{acc}}'(p)).
\]

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

We recorded data in an office room and evaluated the effectiveness of the proposed method by two experiments. The first experiment compared the effectiveness of the three CSP window types. The second experiment investigated the performance of localization of a speaker in an environment with large disturbance noise.

3.1 Recording Conditions

Figure 6 shows the layout of microphones in an experimental environment. The number of the microphones is 16 in our distributed microphone system which is installed in a 4×4 lattice condition under the ceiling, as shown in Fig. 7. The distance between the microphones was 135 cm, and the height of the microphones was 233 cm. As shown in Fig. 6, several noise sources such as a server and workstations existed in the experimental environment. Room reverberation (\( T_{60} \)) was 0.4 sec and ambient noise level was 48.2–56.0 dBA. Thus, this room is a highly noisy environment.

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. A sound source is localized under a condition that the height of the source is given. We investigated the localization accuracy of sound sources for 2-dimensional lattices of hypothetical sound sources; 7.5 cm (3481 point), as shown in Fig. 8.

3.2 Evaluation Experiment of the CSP Window Type

3.2.1 Experimental Conditions

Figure 9 shows the layout of sound sources in an experi-
mental environment. Six different positions indicated by small black circles in Fig. 9 are evaluated as sound source positions. Two loudspeakers are put in two positions which are selected among the six sound source positions. Table 1 shows the concrete position arrangements of the loudspeaker for two sound sources. The position arrangements of the loudspeaker have six patterns from A1 to A6. The reason to choose the combinations of two speakers as shown in Table 1 is that it was assumed that two attendances arbitrarily selected utter at the same time when several persons have a table meeting. Speech materials for sound sources consist of four Japanese sentences spoken by a male speaker and a female speaker, and they are played through loudspeakers and recorded by the distributed microphone system. The direction of the loudspeakers was set to one of four directions; north, east, south, and west. The angle of the loudspeakers was set to the horizontal direction. The height of the loudspeaker was 108 cm. Thus, changing the setting of the loudspeaker, we recorded the data of 96 sentences in total.

3.2.2 Experimental Results

In order to set up the suitable CSP window, the localization accuracy of sound sources was investigated changing the CSP window by three methods; maxCSP, avgCSP, and WavgCSP. The SS-models for three CSP window types and various window widths are obtained with 128 training data of a single sound source. The training data are male speech recordings emitted from a loudspeaker and then recorded by the distributed microphone system. Figure 10 and Fig. 11 show the experimental results of localizing two sound sources. Figure 10 shows the average correct estimation rate of two sound sources, in which a correct estimation is defined such that the distance between the estimated and the correct positions is less than 20 cm. The distance of 20 cm is chosen to evaluate whether estimation is correct or not since two neighboring people, as two sound sources, are assumed to be separated at least by distance of 20 cm due to their physical bodies’ widths. Figure 11 shows the average error distances between the correct and the estimated sound sources. Figure 10 (b) and Fig. 11 (b) show the results of the proposed method. Figure 10 (a) and Fig. 11 (a) show the results of the baseline method which identifies the second largest peak in the original accumulated correlation distribution as the second sound source.

The CSP window is used to consider measurement errors of the microphone positions and it is effective to sound source localization. The localization accuracy of sound sources was maximized by the avgCSP. In Fig. 10 (b), when the CSP window, \( w \), is increased, localization accuracy is
improved. Too large CSP window, $w$, gradually decreases the performance of the sound source localization. However, the CSP window width is not more important than the CSP window type. Even if the value of the CSP window width changes on the appropriate CSP window type, the localization accuracy does not decrease severely. Thus, in the case of an appropriate CSP type, the broad range of the CSP window width gives high localization accuracy.

The localization accuracy of two sound sources is 99.7% for optimized parameters, as shown in Fig. 10 (b). In Fig. 11 (b), average error distances of two sound sources are less than 13.7 cm. Even if a sound source was incorrectly localized, estimated positions were in the vicinity of the correct position. The proposed method clearly has better results than the baseline method. The proposed method subtracts the accumulated correlation distribution around the first sound source from the observed distribution of the accumulated correlation and the subtraction of the accumulated correlation is effective to find the second sound source. These results show that the proposed method accurately estimates two sound sources.

3.3 Localization of a Speaker in an Environment with Large Disturbance Noise

The location of a speaker must be often identified in large noises, such as background music. We evaluated the proposed method supposing that a speaker utters under a disturbance noise with large amplitude. There are two sound sources, and the first and second sound sources are the disturbance noise and speech, respectively. In realistic situations, it is difficult to know a priori the position and the type of the first sound source. It should be discussed how the SS-model is trained for the proposed method.

3.3.1 Experimental Conditions

We tried two types of the first sound source that is a large disturbance noise: a telephone ring emitted from a genuine telephone and music played from a loudspeaker. One audio file of RWC Music Database [12] is used for the music data. The second sound source is a Japanese sentence spoken by a male speaker, and it is played through a mouth simulator.

Figure 12 shows the layout of sound sources in the experiments. The experiments were conducted in two patterns. Table 2 shows the concrete position arrangements of the first and the second sound sources. In pattern 1, the first sound source was fixed at position K and the second sound source was placed on one of nine positions. In pattern 2, the second sound source was fixed at position B and the first sound source was placed on one of nine positions. The SS-model that represents a typical pattern of the accumulated correlation for the first sound source is trained with a different type of the disturbance noise of either music or a telephone ring. In these experiments, the SS-model is trained with single recording of the sound source to investigate the dependency of the SS-model on the recording position. The training data for the SS-model was a priori recorded at one of nine positions D to L.

The directions of the loudspeaker and the mouse simulator were set to the south, and the angle of the loudspeaker and the mouse simulator were set to the horizontal direction. The heights of the mouth simulator and the loudspeaker were 108 cm. Thus, by changing the positions of the first and the second sound sources, and the type of a disturbance noise, 36 different data are recorded. The CSP window type is avgCSP, and the CSP window width is 250 microseconds.
3.3.2 Experimental Results

Figure 13 shows the experimental results of localizing a disturbance noise and a speaker for various arrangements of the first and the second sound sources, shown in Table 2. Figure 13 (a) shows localizing errors for various positions of the second sound source for the pattern 1. Each bar indicates the average error distances between the correct and the estimated sound sources for 8 recording positions of SS-model training data and two types of the disturbance noise. Position K is excluded for the evaluation because it is the same as the position of the first sound source. Figure 13 (b) shows localizing errors for various positions of the first sound source for the pattern 2. Each bar also indicates the average error distances. A position that is the same as the position of the first sound source is excluded for the evaluation. For example, SS-model training data recorded at position D is excluded from the evaluation at position D. Figure 13 shows that the proposed method accurately localizes two sound sources. In addition, the proposed method gives stable performance regardless of the positions of the sound sources.

Figure 14 shows the error distances for various recording positions of SS-model training data while the first and the second sound sources are fixed at position K and A, respectively. Each bar indicates the average error distances for two types of the disturbance noise. In this Figure, first 8 positions from D to L give good performance comparable with the result on position K that matches with the position of the first sound source. Almost same characteristics of stability for the SS-model training can be found for other positions of the first and the second sound sources, shown in Table 2.

This result shows that the proposed method can localize a speaker under a large disturbance noise environment even if the SS-model can not be trained with the data which exactly matches with the large noise as the first sound source.

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 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 SS-model, which is a typical pattern of the accumulated correlation for a single sound source, is subtracted from the observed distribution of the accumulated correlation, and the second sound source is searched again. The experiments were carried out to evaluate the proposed method in a real environment. Our office room is not dedicated to the experiment of sound source localization, thus approximates a real environment. However, the proposed method may not be suitable to all real environments. This method should be further examined to see its capability in adapting to a different environment condition.

As a result of evaluation experiments, we confirmed that the multiple sound source localization estimation performance of the proposed method is superior. In addition, these results show that the proposed method is robust to the training data for the SS-model for different positions in the same room condition. However, this method needs to be examined whether SS-model is also effective in different room conditions, e.g., room size, room reverberation, ambient noise.

In the future, we will investigate the subtraction of the accumulated correlation in order to improve the localization error by the reflection sound. The accumulated correlation is improved 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.
Acknowledgments

This research utilized the RWC Music Database (Popular Music). We would like to thank Mr. Duy Khanh Ninh, Mr. Masato Nakayama, and Dr. Yuki Denda for their assistance during this work.

References


Kook Cho was born in Busan, Republic of Korea, in 1972. He received the B.E., M.E., and Ph.D degrees from Ritsumeikan University in 1998, 2000, and 2005, respectively. From 2005 to 2006, he was a Post-Doctoral Research Fellow at Ritsumeikan University. He is currently a research associate at Ritsumeikan University. His research interests include speech recognition, spoken dialog processing, and acoustic sound signal processing. He is a member of the Acoustical Society of Japan and the International Speech Communication Association.

Hajime Okumura was born in 1982. He received the B.E. and M.E. degrees from Ritsumeikan University in 2006 and 2008, respectively. He currently works for the KYOCERA Corporation.

Takanobu Nishiura was born in 1974. He received the B.E. degree from Nara National College of Technology in 1997 and the M.E. and Ph.D degrees from Nara Institute of Science and Technology (NAIST) in 1999 and 2001, respectively. He received the TELECOM System Technology Award for Students from the Telecommunications Advancement Foundation (TAF) in 2000. From 2001 to 2004, he was a research associate at Wakayama University. He is currently an associate professor at Ritsumeikan University. His current research interests include acoustic sound signal sensors using microphone arrays. He is a member of the Acoustical Society of Japan, Information Processing Society of Japan, Institute of Noise Control Engineering of Japan, Virtual Reality Society of Japan, and International Speech Communication Association.

Yoichi Yamashita was born in Ehime, Japan, in 1959. He received the B.E., M.E., and Dr.Eng. degrees from Osaka University in 1982, 1984, and 1993, respectively. He has worked for the Institute of Scientific and Industrial Research of Osaka University as a Technical Official, a Research Associate, and an Assistant Professor from 1984 to 1997. In 1997, he joined Ritsumeikan University as an Associate Professor in the College of Science and Engineering. He is currently a Professor in the College of Information Science and Engineering. His research interests include speech understanding, speech synthesis, and spoken dialog processing. He is a member of the Acoustical Society of Japan, the Information Processing Society of Japan, the Japanese Society for Artificial Intelligence, the International Speech Communication Association, and IEEE.