Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

Localization of multiple sound sources based on subtraction of accumulated inter-channel correlation

Kook Cho, Takanobu Nishiura and Yoichi Yamashita

College of Information Science and Engineering, Ritsumeikan University, Kusatsu-shi, Japan

PACS: 43.66.Qp

ABSTRACT

The sound source localization plays an important role for extracting the target sound. In this paper we describe the localization of multiple sound sources using a distributed microphone system that is a recording system with multiple microphones dispersed to a wide area. Our algorithm 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 localized by finding the maximum correlation 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 average error distances of the multiple sound sources are less than 13.7cm. Our localization algorithm could realize the multiple sound source localization robustly and stably.

INTRODUCTION

In case of applying a hands-free speech interface to speech recognition, it is very important to localize a sound source accurately. In real environments, ambient noise causes severe performance degradation of speech recognition systems. Precise sound source localization is necessary to mitigate ambient noise for hands-free speech recognition systems.

Conventional methods of estimating direction of arrival (DOA) can be divided to two classes. One class of DOA estimation methods use steering vectors [1]. To search for a source's direction, these methods compare the steering vector of each direction with the spatial correlation matrix and select peaks of the spatial spectrum as estimation results. Therefore, the computational cost depends on the resolution of directions, and high-resolution DOA estimation requires a high computational cost. Another class of DOA estimation methods estimates the direction of a source directly using the phase difference between microphones. The generalized cross-correlation (GCC) based methods [2, 3] belong to this class. A sound source can be theoretically localized using two sets of microphone array by combining two independent directions. When it is anechoic environments, accurate DOA estimation can be achieved by conventional methods. However, in ambient noises and reverberant environments, the DOA estimation performance by conventional methods is greatly deteriorated because reverberation causes an increase in the variance of the phase difference between microphones.

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. Our localization algorithm [4, 5] is a method of multiple sound source localization using a distributed microphone system that is widely distributed and placed under the ceiling of a room. The accumulated correlation algorithm 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 paper is organized as follows. Section 2 describes our localization algorithm based on the inter-channel correlation. Section 3 describes the sound source localization estimations that were carried out in an office room. Section 4 briefly summarizes the conclusions we reached and future work.

SOUND SOURCE LOCALIZATION METHOD

Estimation of TDOA with the CSP method

The direction of the sound source can be obtained by estimating a time delay of arrival (TDOA) between two microphone outputs. The cross-power spectrum phase (CSP) coefficients are calculated by the following equation.

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

$$\tau = \arg\max_k (CSP_{ij}(k)), \quad (2)$$

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, DFT[·] is the discrete Fourier transform, IDFT[·] is the inverse discrete

Fourier transform, the symbol * is the complex conjugate, $CSP_{ij}(k)$ is the CSP coefficients, and τ is an estimated TDOA. The TDOA can be estimated by finding the maximum value of the CSP coefficients.

Sound sources localization based on accumulated the inter-channel correlation

This paper describes the localization algorithm for multiple sound sources based on the inter-channel correlation calculated by the CSP method.

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.(3).

$$k_{ijp} = \frac{\left|\boldsymbol{m}_i - \boldsymbol{s}_p\right| - \left|\boldsymbol{m}_j - \boldsymbol{s}_p\right|}{c} , \tag{3}$$

where $\boldsymbol{m}_i$ is the position coordinate of the i-th microphone, $\boldsymbol{s}_p$(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, $CSPacc(p)$, is derived from Eq.(4) and Eqs.(5), (6), (7).

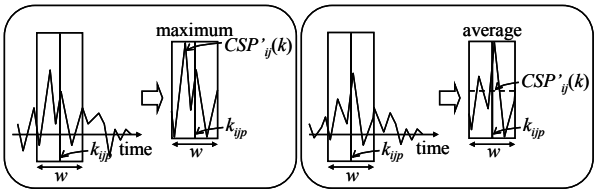
$$CSP_{acc}(p) = \sum_{(i,j)\in S} CSP'_{ij}(k_{ijp}) , \tag{4}$$

$$CSP'_{ij}(k) = \max[CSP_{ij}(k-w), \cdots, CSP_{ij}(k+(w-1)), CSP_{ij}(k+w)] , \tag{5}$$

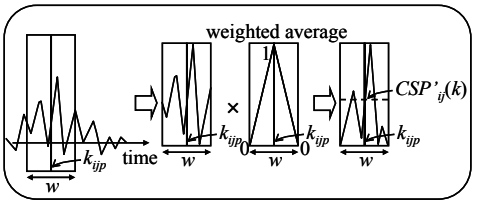
$$CSP'_{ij}(k) = \frac{\sum_{t=-w}^{w} CSP_{ij}(k+t)}{2w+1} , \tag{6}$$

$$CSP'_{ij}(k) = \frac{\sum_{t=-w}^{w} \frac{w-|t|}{w} CSP_{ij}(k+t)}{2w+1} , \tag{7}$$

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.(3). $CSP'_{ij}(k_{ijp})$ 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.(5), (6), and (7). Fig.1 shows these adjusted CSP correlations schematically. Eqs.(5), (6), and (7) are called the maxCSP, avgCSP, and WavgCSP, shown in Fig.1 (a), (b), and (c), respectively. The parameter, w, is a CSP window width that controls a search range in time domain. The CSP window is defined as $|k_{ijp}-t| \le w$. The sound source positions can be estimated by finding the maximum values of the



(a) Maximum CSP coefficient. (b) Average CSP coefficient.



(c) Weighted average CSP coefficient.

Figure 1. Adjusted CSP correlations.

accumulated CSP coefficients by Eq.(8).

$$\hat{l} = \arg\max_{p}(CSP_{acc}(p)) , \tag{8}$$

where $\hat{l}$ is the estimated position of sound source.

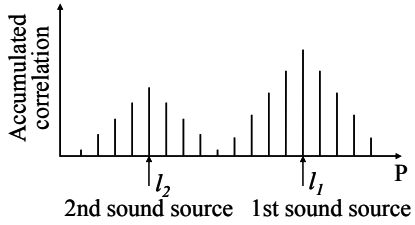
Multiple sound source localization

The procedure for multiple sound source localization is as follows:

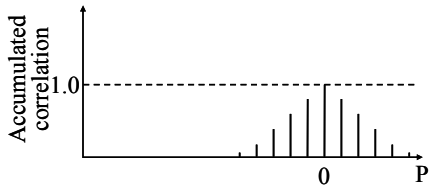
1. After the estimation of the first sound source, a typical pattern of the accumulated correlation for a single sound source is subtracted from the observed distribution of the accumulated correlation.
2. The second sound source is localized by finding the maximum correlation again.
3. The position of multiple sound sources can be estimated by applying the proposed method repeatedly.

Fig.2 illustrates examples of accumulated correlation distribution. In Fig.2 (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 schematical 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, $CSP_{acc}(p)$, shown in Fig.2 (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.2 (b). Fig.2 (c) shows an example of the modified distribution of the accumulated correlation.

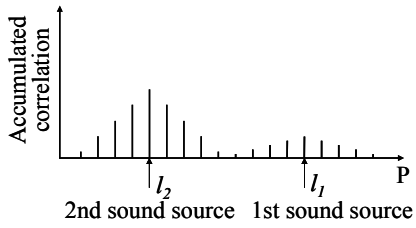
Fig.3 illustrates examples of the SS-model of the accumulated correlation that is obtained with training data. In these figures, the horizontal and vertical axes have the same meanings as those in Fig.2. The SS-model is obtained with the various accumulated correlation distributions for a single sound source shown in Fig.3 (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.3 (b). The accumulated correlation distribution for multiple sound sources is modified by the



(a) An example of observed distribution of the accumulated correlation for two sound sources.

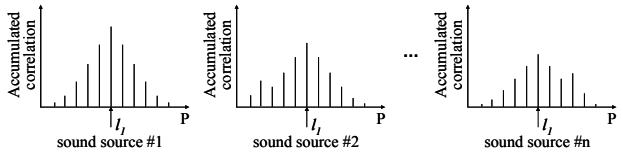


(b) The average distribution of the accumulated correlation that was obtained with accumulated correlation distribution for a single sound source.

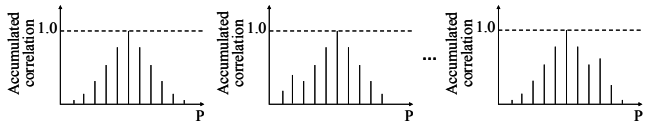


(c) An example of reformed distribution of the accumulated correlation for two sound sources.

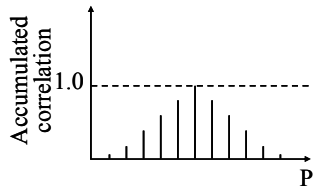
Figure 2. Examples of accumulated correlation distribution.



(a) A various accumulated correlation distributions for a single sound source.



(b) The normalization of the various accumulated correlation distributions.



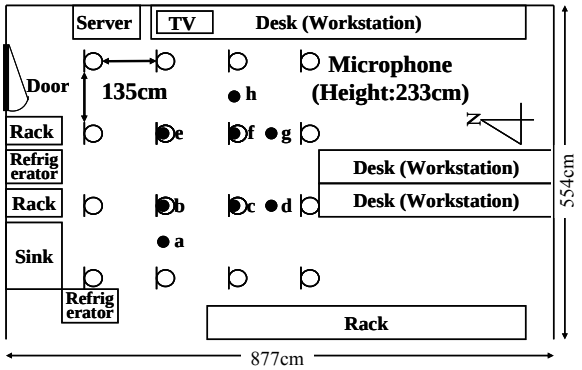
(c) The SS-model is obtained with averaging the various accumulated correlation distributions.

Figure 3. Single Source model (SS-model).

subtraction of the SS-model shown in Fig.3 (c). The modified distribution, $CSP'_{acc}(p)$, is calculated by

$$CSP'_{acc}(p) = CSP_{acc}(p) - CSP_{acc}(\hat{l}_1)Peak(p-\hat{l}_1) \quad (9)$$

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



● : Sound source position (a, b, …, g, h)

Figure 4. An experimental environment.

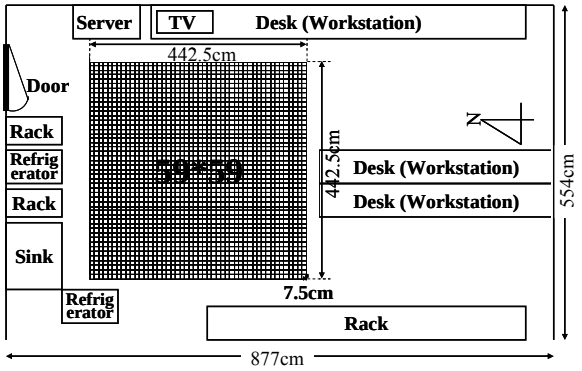


Figure 5. Hypothetical sound sources.

the subtraction of the SS- model. The estimated position of the second sound source, $\hat{l}_2$, is obtained by

$$\hat{l}_2 = \arg\max_p (CSP'_{acc}(p)) \quad . \quad (10)$$

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.

EXPERIMENTAL EVALUATION

Experimental conditions

We recorded data in an actual office room and evaluated the effectiveness of the localization algorithm. Fig.4 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 4×4 lattice condition under the ceiling. The distance between the microphones was 135cm, and the height of the microphones was 233cm. As shown in Fig.4, several noise sources such as a server and workstations existed in the experimental environment. Room reverberation ($T_{[60]}$) was 0.4sec and ambient noise level was 48.2~56.0dBA. Thus, this room is a highly noisy environment.

The sampling frequency was 16kHz, and quantization was 16bits. We tried to evaluate the proposed method with 1,024msec frame length. Sound source localization estimation is conducted for speech periods with 100msec shift interval. The height of the loudspeaker was 108cm. A sound source is localized under a condition that the height of the source is given. We investigated the accuracy of sound source localization for two-dimensional 59×59 point lattice of hypothetical sound sources with the distance between two adjacent points of 7.5cm, as shown in Fig.5.

The number of the sound sources is one or two, in this experiment. In the case of one sound source, speech materials consist of two Japanese sentences spoken by a male speaker, in the case of two sound sources, speech materials 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. 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 Fig.4 are evaluated as sound source positions. In the case of one sound source, one loudspeaker is put in one position which is selected among the eight sound source positions, and in the case of two sound sources, two loudspeakers are put in two positions which are selected among the eight sound source positions. Thus, changing the setting of the loudspeaker, we recorded the data of 128 sentences in the case of one sound source and recorded the data of 96 sentences in the case of two sound sources.

Results for single sound source

The performance of the proposed method is controlled by the CSP window, *w*. We investigated the performance of the CSP window by three methods; maxCSP, avgCSP, and WavgCSP which are defined by Eqs.(5), (6), and (7). Fig.6 shows the average error distances between the correct and the estimated sound sources.

The accuracy in the single sound source localization was maximized by the maxCSP. In Fig.6, average error distances of the single sound source are less than 7.8cm. Even if a sound source was incorrectly localized, estimated positions were in the vicinity of the correct position.

Results for multiple sound sources

In order to set up the suitable CSP window, the accuracy of multiple sound source localization was investigated changing the CSP window by three methods; maxCSP, avgCSP, and WavgCSP. Fig.7 shows the average error distances between the correct and the estimated sound sources.

In Fig.7, 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 the avgCSP. In Fig.7, average error distances of the multiple sound sources are less than 13.7cm. Even if a sound source was incorrectly localized, estimated positions were in the vicinity of the correct position. These results show that the proposed method accurately estimates the multiple sound sources.

Effects of the CSP window

The CSP window is used to consider measurement errors of the microphone positions and it is effective to sound source localization. The maxCSP has good results in the case of the single sound source localization and the avgCSP has good results in the case of the multiple sound source localization. When localizing multiple sound sources, the cross-correlation generates many peaks not only at correct positions but also at incorrect positions in the CSP correlation distribution. In addition, it also reduces the peak at the correct position. Hence, when localizing multiple sound sources, the avgCSP with the average of the CSP coefficient was better than the maxCSP. The sound source localization performance improves by setting the suitable CSP window for the number of the sound sources.

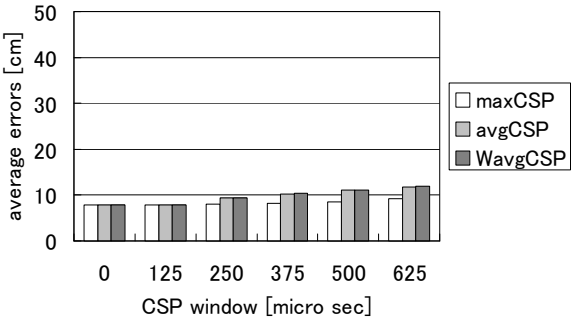


Figure 6. Average errors of the estimated single sound source.

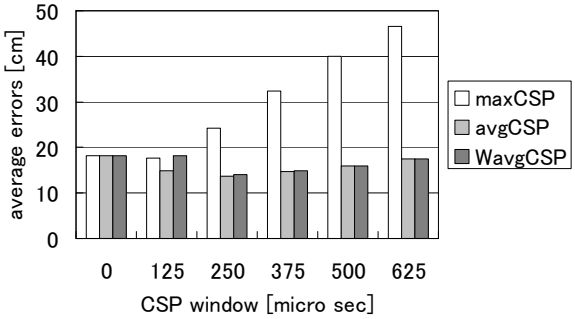


Figure 7. Average errors of the estimated two sound sources.

CONCLUSIONS

This paper evaluated the 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 addition, the performance of the proposed method is improved by the CSP window. The sound source localization performance improves by setting the suitable CSP window for the number of the sound sources.

In the future, we will investigate the subtraction of the accumulated correlation in order to improve the localization error by the reflection sound. In addition, we will attempt to conduct sound source localization with low computational cost.

REFERENCES

1 R.O. Schmidt, “Multiple emitter location and signal parameter estimation,” IEEE Trans. Antennas Propag., vol. **34**, no. **3**, (1986) pp.276–280

2 C.H. Knapp, et al., “The generalized correlation method for estimation of time delay,” IEEE Trans, Acoust. Speech Signal Process., vol. **ASSP-24**, no. **4**, (Aug. 1976) pp. 320–327

3 M. Omologo, et al., “Acoustic source location in noisy and reverberant environment using CSP analysis,” Proc. ICASSP**96**, vol. **2**, (Atlanta, GA, USA, May 1996) pp.921–924

4 K. Cho, et al., “3-Dimensional sound source localization using a distributed microphones system,” Proc. ICA**2007** (International Congress on Acoustics), CAS-04-007, (Madrid, Spain, Sep. 2007)

5 K. Cho, et al., “Localization of Multiple Sound Sources Based on Inter-Channel Correlation Using a Distributed Microphone System,” Proc. of the Interspeech**2008**, (Brisbane, Australia, Sep. 2008) pp.443–446