An Improvement to Extracting and Modifying the Spatial Information in Audio

Lin-Szu Yang and Chia-Ming Chang
Department of Computer Science and Engineering, Tatung University
Taipei, Taiwan
lynnyang.tw@gmail.com, cmchang@ttu.edu.tw

Abstract

In this paper, the method to extract the spatial information and sound sources from the received signal is proposed. The Cross-power Spectrum Phase (CSP) is used to find the arrival time delay from sources to microphone.

If two microphones are considered to locate at two foci, then the trajectory of the difference of time delay found by microphone pair is a hyperbola. Treating a source location is a kind of projection. All of hyperbolas produced by microphone pairs have been collected. Therefore, the back-projection method could be used to reconstruct the source location. After all of the source locations is known, the sound projection is used to extract the sources one by one according the the spatial information. Finally, a sound field would be synthesized.

The result of the simulation verifies that the synthesized sound field similar to original ones is synthesized. Furthermore, a new sound field could also be synthesized according to the spatial information with requirement.

1. Introduction

Usually, the best listening condition must conform with original recording condition, such as the channel numbers, the spatial layout of loudspeaker and listener, the arrangement of loudspeakers, and etc. Reproducing sound field under the unconformable condition, an inappropriate sound field will be obtained. If the spacial information and sound sources in original signal can be extracted, the reproducing sound filed could be adjusted to match the recording condition. Furthermore, a new sound field could also be synthesized with the modified spatial information of sound source.

In Tseng and Chang's research, the spacial information are not easy to extracted from the stereo audio signals. The sources in stereo signal are also difficult to separated [1]. In this paper, some method is applied to improve the performance of Tseng and Chang's research. The improved method is that the spatial information of the sound field is extracted from the received signal by using cross-power spectrum phase (CSP) method and the arrival time delay differences from sources to microphones are gotten. According to the spatial information, each signal at specified location could be extracted, individually. The original sound field would be synthesized in accordance with listening condition.

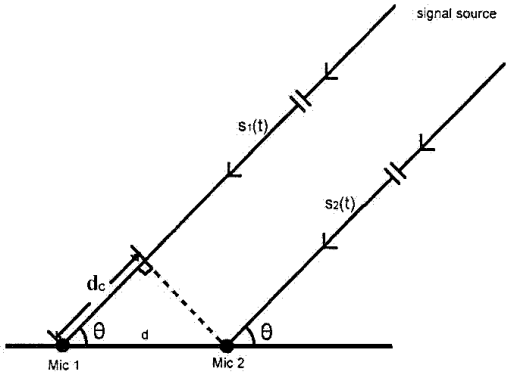

Figure 1: Delay of arrival model [3].

2. Background

There are many researches similar to get the spatial information of the sound signal such as source location [2], direction of arrival (DOA), time delay of arrival (TDOA) [3] and etc. The multiple signal classification (MUSIC) is usually discussed under the condition that the number of microphones is greater than sources and the signal of sources is uncorrelated [4]. Based on given prior information such as signal model, source number, or azimuth can be obtained.

2.1. Delay of Arrival, DOA

The direction of source can be obtained by estimating the DOA between two microphones' outputs. The DOA method is illustrated in Fig. 1. Assume that the sound wave come from infinity, and $M_1(t)$ and $M_2(t)$

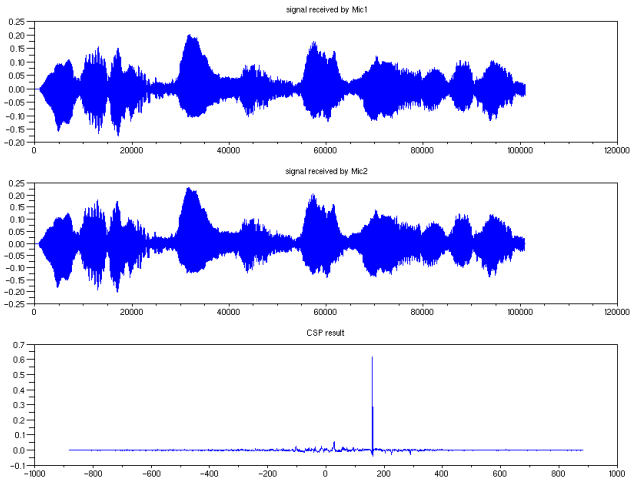

Figure 2: The cross-power spectrum phase of two signals. The last plot is the time difference derived by CPS from two signals plotted above.

are the signals received by Mic1 and Mic2, respectively. The t_1 and t_2 are the signal arrival time to Mic1 and Mic2. The difference of time delay τ is defined as

$$\tau = |t_2 - t_1| \tag{1}$$

Then, the difference of distance d_c could be derived by

$$d_c = \tau * v, \tag{2}$$

where v is the speed of sound in air. Therefore, the direction of sound source could be obtained from the following equation

$$\theta = \cos^{-1}\left(\frac{d_c}{d}\right). \tag{3}$$

2.2. Cross-power Spectrum Phase method

The Cross-power Spectrum Phase (CSP) method is wildly used for finding time delay [5–7]. The phase delay is evaluated from a microphone pair with the cross-correlation function,

$$csp_{ij}(k) = DFT^{-1}\left[\frac{DFT\left[M_i(n)\right] DFT\left[M_j(n)\right]^*}{|DFT\left[M_i(n)\right]| \, DFT\,|[M_j(n)]|}\right], \tag{4}$$

where n and k are the time index, $DFT[\cdot]$ and $DFT^{-1}[\cdot]$ represent the forward and inverse discrete Fourier transform, respectively, and $*$ denote the complex conjugate. The result of CPS method is shown in Fig. 2.

If there is only one sound source, time delay can be estimated by finding the index with the maximum CSP value.

$$\tau = \arg max\left(csp_{ij}(k)\right) \tag{5}$$

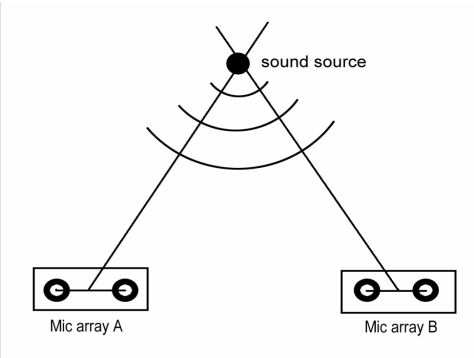

Figure 3: Sound source located with two microphone pairs.

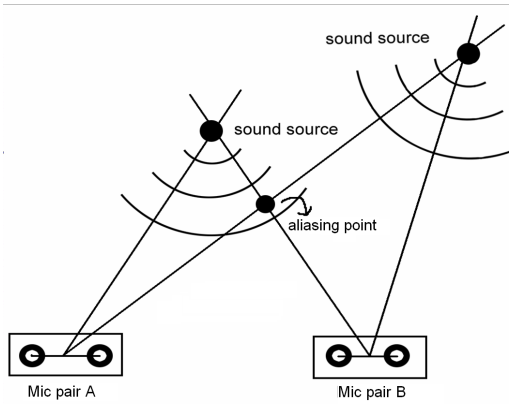

Figure 4: Location problem in the situation of multiple sound sources.

2.3. Location of sound sources

With two microphones, DOA and CSP methods can only detect the direction of source. The sound source can be localized by finding the intersection of two lines obtained from different microphone pairs as shown in Fig. 3.

However, when multiple sound sources need to locate, a problem occurs that the location accuracy has degraded due to the error of cross-correlation among different pairs of sound sources. Fig. 4 illustrates that there is an undesired aliasing point when multiple sound sources are locating.

Therefore, it is necessary to remove the alias point in the situation of multiple sound sources. In next section, we will propose an algorithm to solve this problem.

2.4. Sound Extraction

The method used to extract signal received by microphone array in this paper is the sound projection [8]. It is considered that the sound signal received by a microphone is a kind of projection, a non-straight projection as shown in Fig. 5 In the figure, the relation between distance from microphone and received time is illus-

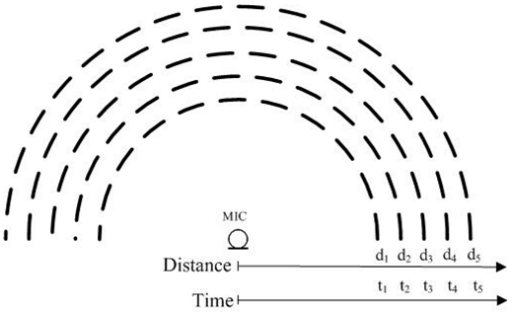

Figure 5: Relation of distance and time from signal to microphone

trated, where d_k is the radius from center, the position of microphone, and t_k is the time that wave propagated from d_k to microphone.

The back-projection technique is used to reconstruct the sound source signal. The point source broadcast the signal with concentric circles; the signals are received by microphones at the same time if the distances from sources to microphones are the same. In other words, the distances from sources to one microphone are the same then signal from different sources reach to the microphone at the same time.

The received signals at time $T + t_k$ is the summation on radius d_k after multiplying with a constant attenuation. To separate the interested source applying back-projection operation to a mixed signal received by the microphone array can be accomplished.

$$s(T) = \sum_{0 \le k < M} \mathcal{F}^{-1}\{\mathcal{F}\{m_k(T)\}|\omega|\}, \tag{6}$$

where T is a time period, $s(T)$ is the source signal, $\mathcal{F}\{\cdot\}$ and $\mathcal{F}^{-1}\{\cdot\}$ are the one-dimensional forward and inverse Fourier transforms, respectively, $m_k(t)$ is the signal received by the kth microphone, $|\omega|$ is the back-projection filter in frequency domain, and M is the size of microphone array.

3. System Description

In the proposed method, the spatial information of sound sources are obtained from the received signal. According to the spatial information, sources at specified location are extracted. A sound field could be synthesized in accordance with the listening condition. The system framework is shown in Fig. 6.

3.1. Sound Receiving

The microphone array has a good quality on acquisition of acoustic signal [9]. The major usages of microphone array in this paper are the sound source location and source signal separation.

The sound signals are received by a microphone array. The received signal $x_j(t)$ at jth microphone consists of M sensors may be modeled as

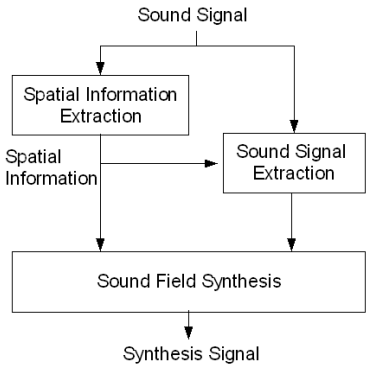

Figure 6: System framework.

$$x_j(t) = \sum_{i=1}^{N} [h_{ij}(t) * s_i(t) + n_{ij}(t)], \tag{7}$$

where $s_i(t)$ is the signal come from the ith source and N is the index ofsource. The $h_{ij}(t)$ is the impulse response of the channel from the ith source to the jth microphone and the $n(t)$ is the additive noise. In a non-reverberation space, the impulse response $h_{ij}(t)$ could be expressed as

$$h_{ij}(t) = \begin{cases} a(k_{ij}), & t = k_{ij} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

$$a(k_{ij}) = \alpha * d_{ij}^{-2}, \tag{9}$$

where k_{ij} is the propagation time from ith source to jth microphone, d_{ij} is the distance between the ith source to jth microphone, and α is the standard attenuation coefficient.

3.2. Spatial Information Extraction

Locating the sound source can be achieved by estimating the time delay from received signals. Based on previous researches, the signal received by microphone could be treated as the summation of the source spectrogram with shift in time and attenuation. The relative time delays could be retrieved by applying auto-correlation to the spectrogram of received signals as shown in Fig. 7. In practice, time delay could not be evaluated accurately using this method. It is due to the reasons that relative time delay are too close and the measurement of time and frequency resolution of spectrogram is an issue. Therefore, estimating time delay by auto-correlation of spectrogram is difficult.

Signals came from a specific position are received by each microphone in microphone array with different arrival time. That is cause of the difference distances from sources to microphones. The signal come from same source received by different microphone can be assumed that a delayed copy of the source signal. The Cross-power Spectrum Phase (CSP) can be used

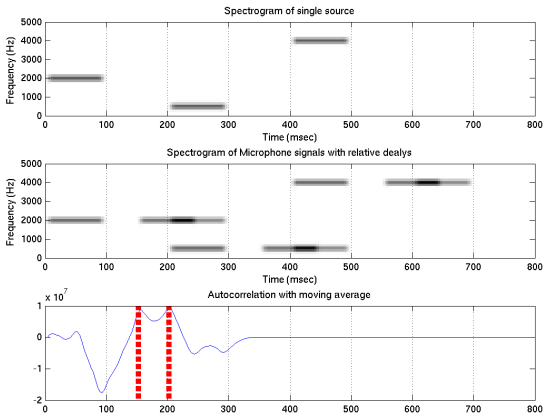

Figure 7: Auto-correlation of spectrogram.

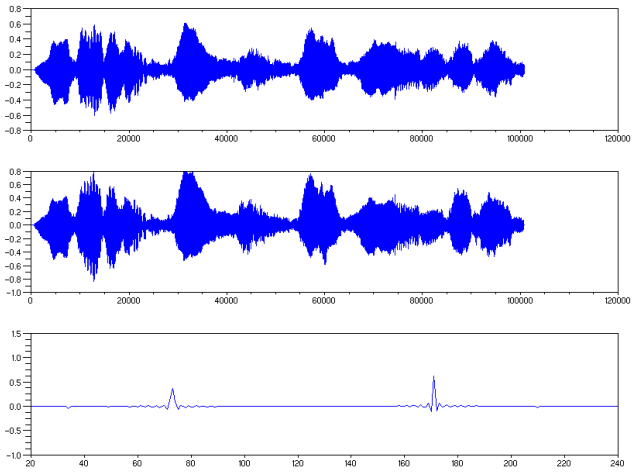

Figure 8: CSP result of multiple sources. The horizontal axes in first two plots are the sampling index, and the horizontal axis in last plot is the difference of indexes.

to find arrival time delay indexed by the peak of the correlation function. The index of the maximum value in CSP results represents the delay time from source to microphone.

We observed that there are more than one peak in the CSP results when there are multiple sources as shown in Fig. 8. They represent the difference of delay time from sources to microphones. After the time delays have been evaluated, the distances difference from each source to microphone pairs could be evaluated by the following equation:

$$\text{distance difference} = \frac{\text{delay time}}{F_s * v} \qquad \text{(meter)} \quad (10)$$

where F_s is the sampling rate and v is the sound speed.

Assume the microphones are located at two foci. We will observe that the trajectory of the distance difference found by microphone pair is a hyperbola as shown in Fig. 9. If the processed signal is stationary over the observation interval then the delay of arrival time

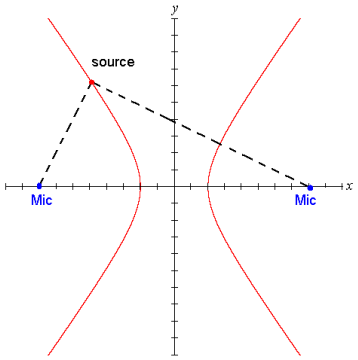

Figure 9: The trajectory of points with constant time delay is a hyperbola.

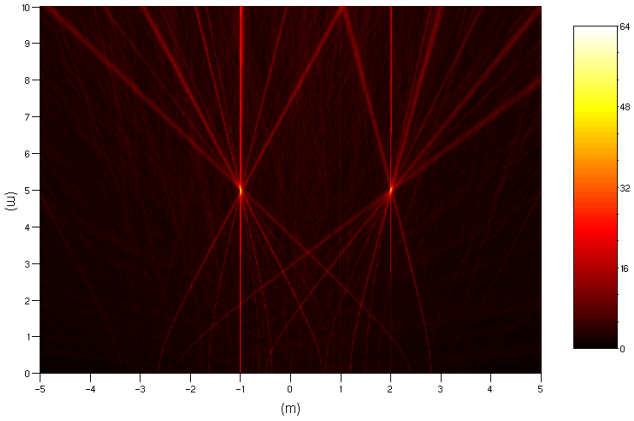

Figure 10: Projection lines produced by microphone pairs. The microphones are set at bottom and the center is indexed as 0.

from sources to microphone pairs will be fixed. Consequently, we could say that the source location might probably located on the curve of a hyperbola obtained from the analysis of microphone pairs.

Finding the source location is a process of non-straight projection. All of hyperbolas produced by each microphone pair have been collected. Therefore, the back-projection method could be used to find the source location. Regions where projection lines from different hyperbola intersect indicate the area of source location as shown in Fig.10.

Gather the high density intersect of projection lines and classify them by using a clustering algorithm DBSCAN. After classification, the centroids of each area would stand for the source location. The centroid is evaluated by the following equation.

$$\text{centroids} = \frac{\sum_{i=1}^{k} (X_i * D_{X_i})}{\sum_{i=1}^{k} X_i}, \qquad (11)$$

where X_i is the probable source location, D_{X_i} is the density of X_i and k is the amount of probable source location in the cluster.

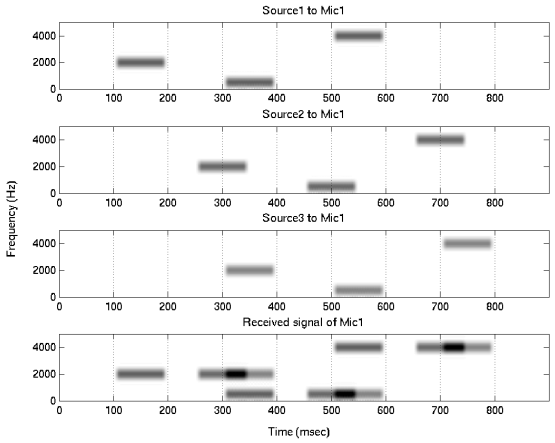

Figure 11: Spectrograms of received signal

3.3. Source Extraction

The recording of microphone is a procedure to sum up sources with different time delays and energy attenuation according to the distances between microphone and sources. In the previous research, we suppose the procedure of receiving signals from sources to microphone could be treated as the procedure of image corrupted by motion blur when regarding each source as the same "time-frequency" image as shown in Fig. 11. Therefore, the image-restoration technique was applied on the sound source separation. But this method is restricted that the sources have to be similar and the source positions have to set aligned in equidistant also.

The sound projection method is used to extract the sources one by one according the spatial information of the sound field. The process of signal received by microphone is taking to be a kind of non-straight projection. The signal received by microphones is treated as the results of projection along the circles with the same center. The back-projection operation is used to reconstruct the specified source signal in our method.

3.4. Sound Field Synthesis

After extraction the spatial information of each sound sources, sound field synthesis is performed. Original sound field would be synthesized based on the spatial information and extracted source signal. A new sound field would also be synthesized with requirement. Modified spatial information can be used to produce a different sound filed. The spatial layout of the sources is changed on demand. It could even take off the sources and join a new one source in the new sound field.

4. Simulation

For limitation of hardware, a software simulation is used to verify the proposed method. Assuming that there is a two-dimensional room with 10 meters width

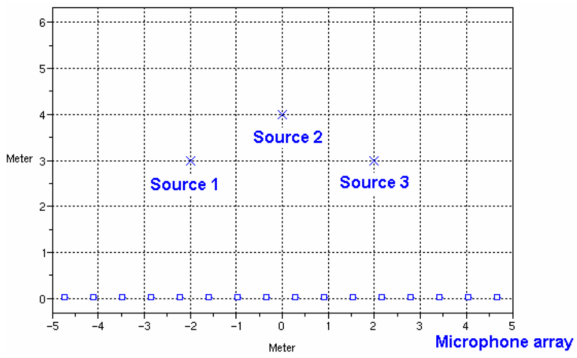

Figure 12: System Arrangement.

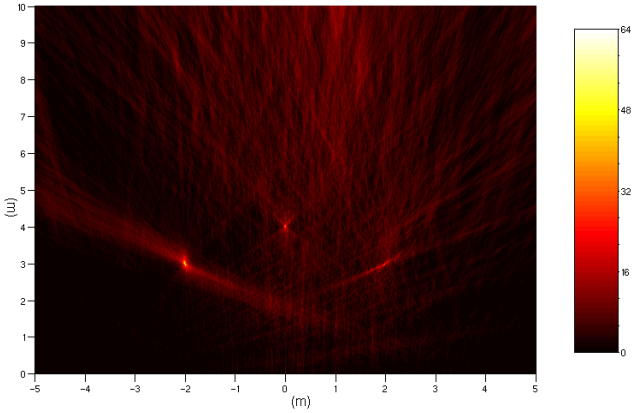

Figure 13: Estimated location of sound sources playing similar music.

and 10 meters length without reverberation. Sixteen microphones are set in the bottom and three sound sources are placed in the locations of (-2,3), (0,4), and (2,3) as shown in Fig. 12.

4.1. Experiment

Two cases are considered in the experiments. In the first case, there are three similar sound sources playing the same melody. In another case, there are three different kinds of instruments playing different music in the same moment.

4.1.1. Similar Source

There are three violins play the same melody in the room. We generate the similar sound sources from one solo source by using of phase shifting and random samples shifting. In such a way, it generates the effect of difference and asynchronous among sources.

By using proposed method, the distribution of similar sound source is demonstrated in Fig. 13. The colors near to white indicate there are more projection lines intersecting. It is useful to represent where the source is. The estimated source location are (-2,3), (0,4), and (2,2.99).

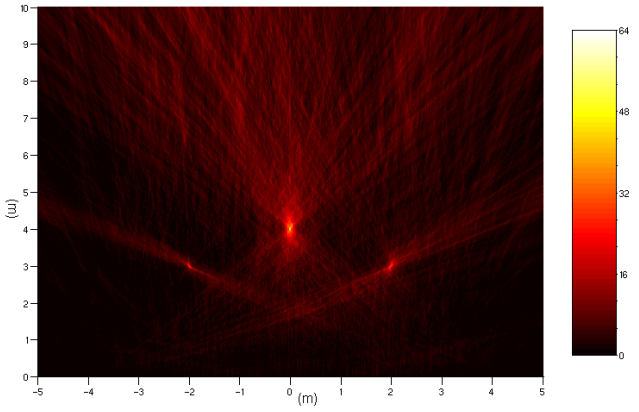

Figure 14: Estimated location of sound sources playing different music.

4.1.2. Mix Sound Sources

Three different sound sources are playing in the room. There are violin, guitar, and piano. And they play different music. Microphones will receive a mix signal. The simulation result of mix sound source location is demonstrated in Fig. 14. Estimated similar source location are (-2,3), (0,4.03), and (1.99,2.99).

4.2. Performance Evaluation

In order to verify the performance of our system, we compute the variance. The variance averages squared distance between original source location and estimated source location. It is a measurement to capture the scale error.

$$\text{mean square error} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{X}_i - X_i\right)^2, \qquad (12)$$

where $\hat{X}$ is the estimated source location, X is the original source location and N is the source number. The Euclidean distance is used when counting the difference of two positions. The mean square error of two experiments are shown as in the Table. 1.

5. Conclusion and Future work

The synthesis solution was proposed by decomposing received signals as spatial function and single sources. The simulation result shows that the method for localization of source is working. Spatial information could be extracted accurately. There are no restrictions on the arrangement of source aligned in line with equidistant. A new sound field can also be synthesized by different spatial information with requirement.

Table 1: The mean square error of source location

	Variance
Similar sources	0.0000333
Mix sources	0.0001667
Average	0.0001505

References

[1] H.-Y. Tseng and C.-M. Chang, "Extracting and modifying the spatial information in stereo audio," in *Proceedings of 2006 International Computer Symposium (ICS 2006)*, Dec. 2006.

[2] T. Nishiura, T. Yamada, S. Nakamura, and K. Shikano, "Localization of multiple sound sources based on CSP analysis with a microphone array," in *Proceedings of 2000 IEEE International Conference on Acoustic, Speech, and Signal Processing*, vol. II, no. 6, 2000, pp. 1053–1056.

[3] J.-T. Chien, J.-R. Lai, and P.-Y. Lai, "Microphone array signal processing for far-talking speech recognition," in *Proceedings of 2001 IEEE Third Workshop on Signal Processing Advances in Wireless Communications (SPAWC '01)*, Mar. 2001, pp. 322–325.

[4] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[5] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," in *Proceedings of 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '94)*, 1994, pp. 273–276.

[6] ——, "Acoustic source location in noisy and reverberant environment using CSP analysis," in *Proceedings of 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96)*, 1996, pp. 921–924.

[7] P. Svaizer, M. Matassoni, and M. Omologo, "Acoustic source location three-dimensional space using crosspower spectrum phase," in *Proceedings of 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, 1997, pp. 231–234.

[8] C.-M. Chang and C.-H. Peng, "Applying the filtered back-projection method to extract signal at specific position," in *Proceedings of 2005 National Computer Symposium (NCS 2005)*, Dec. 2005.

[9] H. F. Silverman, W. R. Patterson, and J. L. Flanagan, "The huge microphone array," *IEEE Concurrency*, vol. 6, no. 4, pp. 36–46, Oct.-Dec. 1998.