# Full Azimuth Multiple Sound Source Localization with 3-Channel Microphone Array

**Article** *in* IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences · April 2012
DOI: 10.1587/transfun.E95.A.745



**4 authors:**


Suwon Shon
Korea University
**24** PUBLICATIONS **37** CITATIONS
SEE PROFILE


David K. Han
**77** PUBLICATIONS **337** CITATIONS
SEE PROFILE


Jounghoon Beh
University of Maryland, College Park
**20** PUBLICATIONS **153** CITATIONS
SEE PROFILE


Hanseok ko
University of Maryland, College Park
**264** PUBLICATIONS **1,049** CITATIONS
SEE PROFILE

PAPER

# Full Azimuth Multiple Sound Source Localization with 3-Channel Microphone Array

**Suwon SHON**[†], **David K. HAN**[††], **Jounghoon BEH**[†††], *Nonmembers*, *and* **Hanseok KO**[†a)], *Member*

**SUMMARY** This paper describes a method for estimating Direction Of Arrival (DOA) of multiple sound sources in full azimuth with three microphones. Estimating DOA with paired microphone arrays creates imaginary sound sources because of time delay of arrival (TDOA) being identical between real and imaginary sources. Imaginary sound sources can create chronic problems in multiple Sound Source Localization (SSL), because they can be localized as real sound sources. Our proposed approach is based on the observation that each microphone array creates imaginary sound sources, but the DOA of imaginary sources may be different depending on the orientation of the paired microphone array. With the fact that a real source would always be localized in the same direction regardless of the array orientation, we can suppress the imaginary sound sources by minimum filtering based on Steered Response Power — Phase Transform (SRP-PHAT) method. A set of experiments conducted in a real noisy environment showed that the proposed method was accurate in localizing multiple sound sources.

***key words:*** *multichannel array processing, SRP-PHAT, sound source localization (SSL)*

## 1. Introduction

When there are multiple sound sources, it is often difficult for an automatic speech recognition system to separate and accurately recognize the speech content of the sound source of interest. To achieve better recognition accuracies in such cases, preprocessing of acoustic signals to separate and accurately localize each sound source is necessary. One of the methods of Sound Source Localization (SSL) is by a microphone array. In this approach, the direction of a sound is found by considering the time delay of arrival (TDOA) [1] among the microphones in the array with the associated array geometry. SSL has been an active research topic for the last several decades. An active area within the topic is to solve the localization problem of multiple sources in a noisy environment with a minimal number of microphones.

To obtain multiple DOAs of multiple sound sources, there are several key approaches proposed including MUltiple SIgnal Classification (MUSIC) method [2], phase difference method [3], [4], Steered Response Power — Phase Transform (SRP-PHAT) method [5].

Manuscript received September 11, 2011.
Manuscript revised November 21, 2011.

[†]The authors are with School of Electrical Engineering, Korea University, Seoul, Korea.
[††]The author is with Office of Naval Research, Arlington, VA, USA.
[†††]The author is with University of Maryland College Park, MD, USA.
a) E-mail: hsko@korea.ac.kr

DOI: 10.1587/transfun.E95.A.745

The MUSIC method employs an eigenvector decomposition of the incoming signals acquired by the array. Its ability of separating sound sources is limited by the number of microphones used wherein up to M-1 sources can be localized with M microphones. The phase difference SSL is not limited by the number of microphones. However, the method assumes that the sound sources mutually satisfy W-disjoint orthogonal property under the assumption of the speech sources being sparse in time-frequency domain. But this assumption is no longer valid in noisy environments [4], [6].

The SRP-PHAT method is based on Generalized Cross Correlation — Phase Transform (GCC-PHAT) method, also known as Cross-power Spectrum Phase (CSP) [7]–[10]. SRP-PHAT is performed by finding local maxima of SRP [5], [7], [11]. It is a very popular DOA estimation method and has been shown to perform very well in noisy environment [7], [12]–[15]. However, finding local maxima are affected by noise level. Moreover, if a pair of microphones estimates full azimuth DOA, an imaginary sound source that has the same TDOA as the real source is also generated. This imaginary sound source has greater effect on localizing multiple sound sources than the effect posed by noise. This imaginary sound source problem is usually handled by larger number of microphones to suppress imaginary sources [16]. Some avoided this problem all together by considering only the sources directly in front of the microphone arrays [17]–[19] and others avoided this problem by only considering a single source [20].

In this paper, we tackle the issue of imaginary sources generated from multiple sources in a three channel microphone system. We propose a novel method of suppressing imaginary sources by a minimum search of angular power distributions from the microphone pairs. Our aim is to remove imaginary sources in full azimuth with minimal number of microphones for the purpose of simplified and affordable hardware and reduced order of computations. Throughout this paper, we assume a far field acoustic model, thus acoustic waves are assumed to be planar.

## 2. Problem Description

Consider a pair of microphones in 2-dimensional space. As shown in Fig. 1, an imaginary sound source is perceived by a microphone array because of the same TDOA. $\tau_{lq,\theta s}$ is TDOA between the $l$-th and the $q$-th microphones of a sound source at $\theta_s$ as shown in Eq. (1)

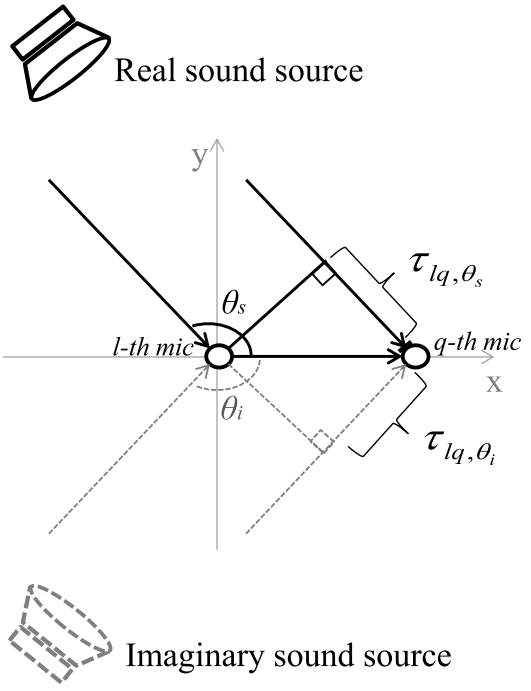



**Fig. 1** Example of real and imaginary sound source.

$$\tau_{lq,\theta_s} = \frac{d\sin(\theta_s)}{c} \tag{1}$$

where $d$ is the distance between the microphones, $c$ is the speed of sound.

Using Eq. (1), the sound source azimuth $\theta$ can be calculated as a function of $\tau_{lq,\theta}$ as

$$\theta = \sin^{-1}\left(\frac{c\tau_{lq,\theta}}{d}\right). \tag{2}$$

Because the inverse cosine is a multi-valued function, Eq. (2) always yields two solutions between $-180^\circ \le \theta \le 180^\circ$. One of these is the azimuth of the real sound source and the other is of the imaginary source.

In our implementation, a pair of microphones is used at a time for SSL. If there are 3 microphones to an SSL, there would be 3 imaginary sound sources because of the 3 paired microphone arrays as in Fig. 2. The source direction, denoted as $\theta_s$, is based on the $x$-$y$ coordinate shown in Fig. 2. Since we assume that the far field model, $\theta_s$ measured in reference to the $x$-$y$ coordinate system shown should be identical to the direction computed by each of the microphone pair. Imaginary sound source 1 is from the microphone pair of 1 and 3, source 2 is from the microphones 2 and 3, and source 3 is from the microphones 1 and 2, respectively.

## 3. Full Azimuth Multiple SSL Method

### 3.1 Estimation of TDOA with SRP-PHAT

A brief description of a conventional SRP-PHAT algorithm is as follows. GCC of the $l$-th and $q$-th microphone signals is

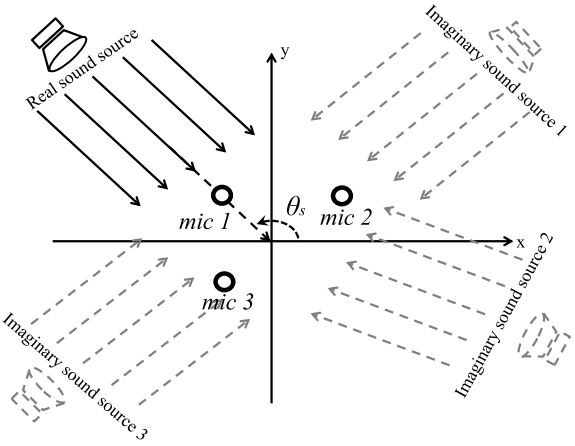



**Fig. 2** Example of imaginary sound sources.

$$R_{lq}(\tau, n) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\Psi_{lq}(\omega, n)X_l(\omega, n)X'_q(\omega, n)e^{j\omega\tau}d\omega \tag{3}$$

where $n$ is the frame index, $X_l(\omega, n)$ the Short-Time Fourier Transform (STFT) of $l$-th microphone, $w$ the frequency index, and $\Psi_{lq}(\omega, n)$ denoting a weight function. Although many different weighting functions can be applied, the Phase Transform (PHAT) has been found to perform quite well under realistic acoustical conditions [8]. The PHAT weight function can be defined as:

$$\Psi_{lq}(\omega, n) \equiv \frac{1}{|X_l(\omega, n)X'_q(\omega, n)|}. \tag{4}$$

Using GCC-PHAT of the $l$-th and the $q$-th microphone signals, we can estimate the TDOA as

$$\hat{\tau}_{lq} = \arg\max_{\tau}(R_{lq}(\tau)) \tag{5}$$

The SRP-PHAT algorithm is summing GCCs. Thus, the SRP can be expressed as

$$P_n(\Delta_1\Delta_2\ldots\Delta_M) = \sum_{l=1}^{M}\sum_{q=l+1}^{M}\int_{-\infty}^{\infty}\Psi_{lq}(\omega, n)X_l(\omega, n)X'_q(\omega, n)e^{j\omega(\Delta_q-\Delta_l)}d\omega. \tag{6}$$

where $\Delta_m$ is the propagation delay from source to the $m$-th microphone. It is guaranteed that the global maximum of the SRP corresponds to the location of a sound source.

### 3.2 SSL

In the far field model, by TDOA of specific direction $\theta$, Eq. (6) becomes

$$P_n(\theta) = \sum_{l=1}^{M}\sum_{q=l+1}^{M}\int_{-\infty}^{\infty}\Psi_{lq}(\omega, n)X_l(\omega, n)X'_q(\omega, n)e^{j\omega\tau_{lq,\theta}}d\omega = 2\pi\sum_{l}^{M}\sum_{q=l+1}^{M}R_{lq}(\tau_{lq,\theta}, n) \tag{7}$$

where $\tau_{lq,\theta}$ can be calculated by Eq. (1). An estimated azimuth of a sound source can be found from the maximum

SRP as

$$\theta_s = \arg \max_{-\pi \le \theta \le \pi} (P_n(\theta)). \quad (8)$$

The conventional SRP-PHAT algorithm sums GCCs of all paired microphone arrays. Hence, the SRP not only contains real sound sources but it also contains imaginary sound sources. Of course, the SRP in the direction of a real sound source would exhibit a high value in every GCC of paired microphones but it would not in the direction of imaginary sources. The azimuth of a real sound source always exists among the two solutions of Eq. (2) independent of the paired microphone array geometry. However, the azimuth of an imaginary sound source changes depending on the geometry of the paired microphone array as shown in Fig. 2. Therefore, the resultant SRP in the direction of an imaginary sound source would be smaller than that of a real sound source. In the case of a single source, discriminating the real source from an imaginary one is trivial, since the direction of the maximum SRP corresponds to the real source. In the case of $M$ multiple sound sources, the approach of finding $M$ largest values and corresponding directions of the SRP may not find directions of real sources since an imaginary source may also yield a large SRP in the summing process. Therefore, imaginary sound sources must be suppressed in the case of multiple SSL.

We propose suppressing imaginary sound sources by minimum filtering method instead of summing all GCC-PHAT of paired microphone arrays. A modified SRP can be expressed as

$$\hat{P}_n(\theta) = 2\pi \cdot \min_{1 \le l,q \le M} \left( R_{lq}(\tau_{lq,\theta}, n) \right). \quad (9)$$

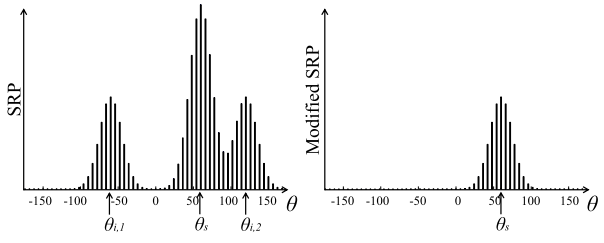
Suppose there is a sound source in 60° detected by a microphone array as shown in Fig. 2. Figure 3 presents examples of the corresponding GCC. Figure 3(a) shows when the microphones 1 and 2 are paired and Fig. 3(b) shows when the microphones 1 and 3 are paired. $\theta_{i,p}$ is the azimuth of an imaginary sound source and $\theta_s$ is the azimuth of a real source of the $p$-th paired microphone array.

Conventional SRP sums GCC, so that the resultant SRP is like Fig. 4(a) shown below. We propose a logical conjunction type filtering approach using a minimum filter. Instead of summing GCC, we apply the minimum rule such that the resultant SRP contains the minimum response from each beam directions obtained from all the microphone pairs. Therefore, the minimum operation would result Fig. 4(b). As it is apparent, the minimum operation successfully suppressed the imaginary sound source, thus providing the direction of the real source. $\hat{P}_n$ would remain approximately the same as $P_n$ in the direction of a real source location.

In the case of single sound source, both the conventional SRP and the proposed method were successful in correctly finding the real source direction. When there are more than one sound source, however, the conventional SRP may not correctly determine the real direction of each sound source. Consider an example of three sound sources located at $-170°(\theta_{i,1})$, $45°(\theta_{i,2})$, and $150°(\theta_{i,3})$ in our coordinate system with the source power ratio at 8:7:3 respectively. Then the sound source directions detected by microphone pairs 1& 2 may look like Fig. 5.

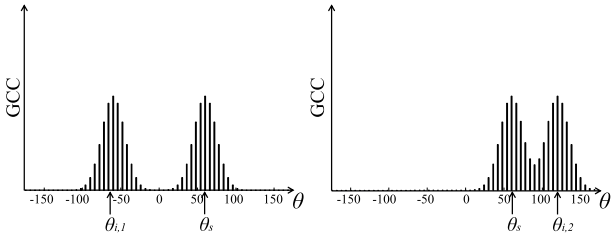
By summing the response from each pair, effective directional response by the conventional algorithm is shown in Fig. 6(a). It is clear that there are more apparent source directions than the real number of sources. The high power sources at $-170°$ and $45°$ match correctly with the real source directions. However, the power level of the real source at 150° does not necessarily stand out over the imaginary sources shown at $-45°$, $-10°$ and $170°$. Due to the summing process in the conventional method, power levels of imaginary sound sources may, at times, equal or exceed that of the real sources. Thus, the conventional method may lead to incorrect directions of the sound sources.

The proposed algorithm based on the minimum filtering results as shown in Fig. 6(b). By the minimum process, the fictitious power associated with an imaginary sound



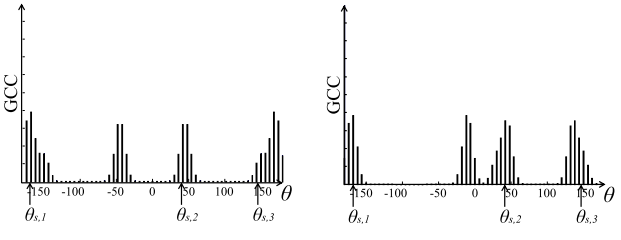

(a) Conventional (b) Proposed

**Fig. 4** Single sound source examples of SRP using two methods (conventional and proposed).



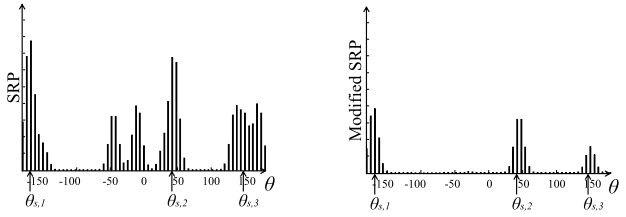

(a) Pair 1 (b) Pair 2

**Fig. 3** Examples of GCCs for a single sound source.




(a) Pair 1 (b) Pair 2

**Fig. 5** Examples of GCCs for multiple sound sources.




(a) Conventional (b) Proposed

**Fig. 6** Multiple sound sources examples of SRP using two methods (conventional and proposed).

source is suppressed correctly. Only the powers of the real sources remain. Although it wasn't quite obvious of the advantage of the proposed method in finding the direction of a single source, as the number of the source increases, the proposed method outperforms in finding real source directions.

## 4. Experiments

### 4.1 Experimental Conditions

Various experiments were conducted in a noisy environment to validate the effectiveness of the proposed method. Figure 7 shows the experimental environment and placement of sound sources and the microphone array.

There are several noise sources like workstations and a server. The room reverberation ($T_{[60]}$) was 0.3 sec and ambient noise level was 53 dBA average. The microphones are arranged as shown in Fig. 7, and the sound sources and the microphones are all placed 100 cm above the floor. The signal level of sound source at the microphone array is 63 dBA average, so the Signal-to-Noise Ratio (SNR) is 10 dB average.

As a first step in the experiment, to determine the 3-channel array performance, we first examine its angular resolution by an experiment. We used two loudspeakers with one fixed at 108° and the other located initially at 30°. The one at 30° was moved closer to the one at 108° until the SRP merged into one peak. For 16cm microphone distance and 3 channel microphone array like Fig. 7, the angular resolution was 15° as in Fig. 8.

There are 6 loudspeakers 100 cm away from the microphone array in different directions.

To fully evaluate robustness of the proposed method, we considered a variety of music types as well as vocals of male and female as shown in Table 1. They are each 5 minutes in duration.

The microphone array recorded with 16 kHz sampling frequency and 16 bit quantization. We performed SSL with 1024 sample frame length and 512 sample shift intervals. The SSL performed about 9000 times in 5 minutes.

A total of 4 different sets of experiments were conducted. In the first, only two sound sources were employed. Next, 3, 4 and 5 sound sources were added progressively as shown in Table 2.

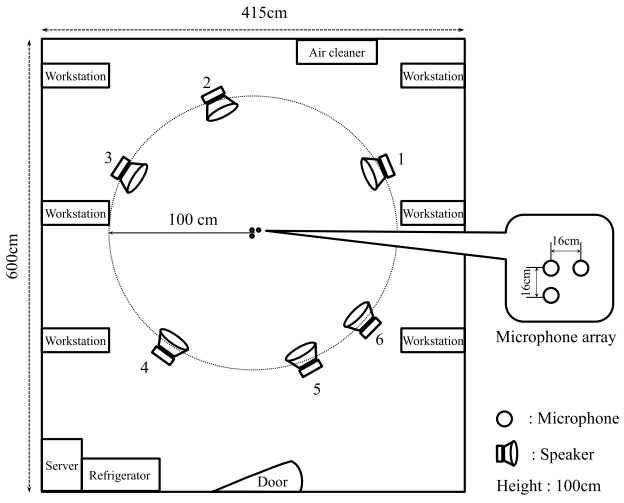
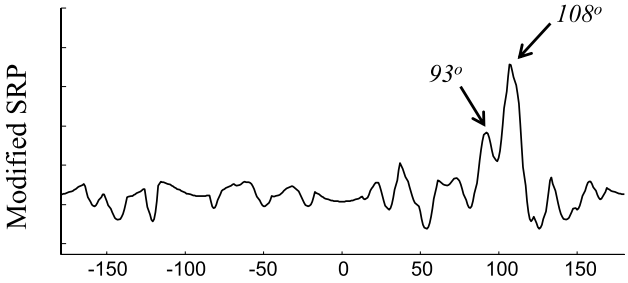
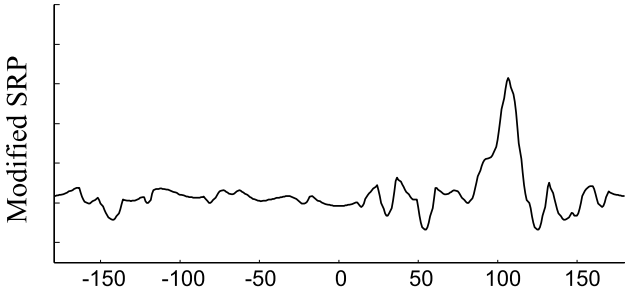



**Fig. 7** Experimental environment for various tests with different sound sources.




(a) SRP example when the angle between two sound sources is 15°




(b) Merged peak example of SRP when the angle between two sound sources is below 15 °

**Fig. 8** Angular resolution test.

### 4.2 Experimental Results

The performance of the proposed method was compared with a conventional SRP-PHAT and Cho's method [7]. For Cho's method, we implemented the most favorable condition for the method's performance. Specifically, we used his weight average CSP method with scaled dimensions of 30 [micro sec] CSP window at 16 cm microphone array dis-

**Table 1** Sound sources information.

| Loudspeaker Index | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Angle | 25 | 120 | 160 | 225 | 290 | 330 |
| Genre | Rock | Hiphop | Classic | Country | Soft music | R&B |
| Vocal | Woman | Men | none | Man | Woman | Man |

**Table 2** Recorded sets.

| Number of sources | Loudspeaker set used |
|---|---|
| **2** | ( 1 , 3 ) , ( 4 , 6 ) , ( 2 , 5 ) |
| **3** | ( 1 , 4 , 6 ) , ( 2 , 3 , 5 ) |
| **4** | ( 1 , 2 , 5 , 6 ) , ( 2 , 3 , 4 , 5 ) |
| **5** | ( 1 , 2 , 3 , 4 , 5 ) , ( 2 , 3 , 4 , 5 , 6 ) |

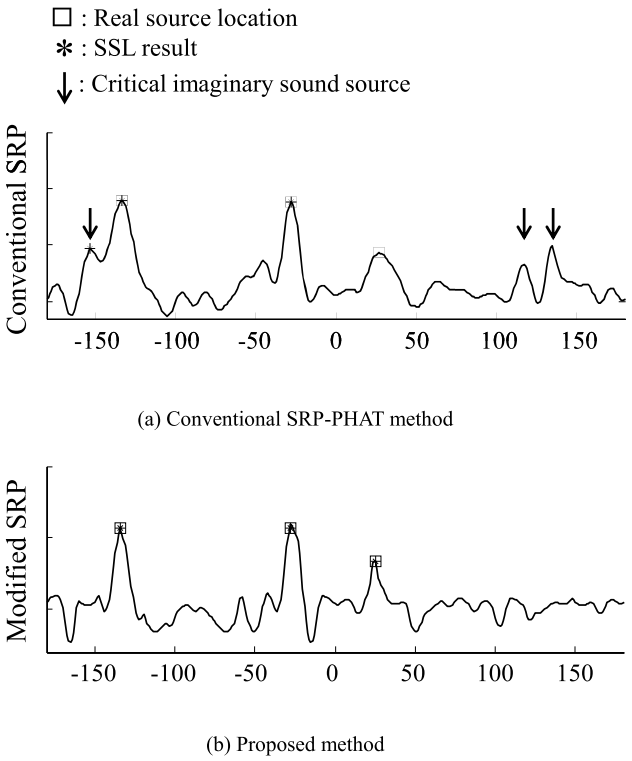
**Table 3** Localization accuracy of multiple SSL according to number of sources.

| Method / Number of sources | Accuracy (%) | | |
|---|---|---|---|
| | Conventional method | Cho's method | Proposed method |
| **2** | 88.96 | 90.62 | 98.41 |
| **3** | 77.94 | 84.80 | 90.15 |
| **4** | 64.93 | 70.26 | 78.33 |
| **5** | 60.72 | 66.91 | 70.26 |

tance (Cho used 250 [micro sec] CSP window at 135 cm), and single source model subtraction for multiple SSL.

Table 3 shows localization accuracy of estimating multiple SSL. Tolerance was ±5° because loudspeaker is 30 cm wide, so it takes up about 10° from the microphone array azimuth. When there are 2 sources the accuracy of proposed method improved by 10% compared to that of the conventional method and 8% compared to that of Cho's method as shown in Table 3.

Figure 9 shows the SSL results when there are 3 sources known. We obtained the averaged version of modified SRP over 140 time frames. In Fig. 9(a), the conventional method localized imaginary sound sources as real sources. This is due to the imaginary sound sources having higher power than real sound sources upon the summing process of SRP-PHAT. Therefore, the SSL failed in finding some of the real sound sources. However Fig. 9(b) indicates that the proposed method suppressed imaginary sound sources successfully. Therefore, the method localized the real sound sources accurately.




(a) Conventional SRP-PHAT method

(b) Proposed method

**Fig. 9** Comparison of SSL results when there are 3 sources known.

When there were sound sources greater than three, the proposed method was yet able to perform better than the other methods. Nevertheless, it was observed that the proposed method also mis-localized real sources in some cases as the number of sound sources increased. This is mainly due to the imaginary sources appearing at the same angular direction in the SRP over more than one array pair. In such a case, the minimum filtering proposed here would not suppress the GCC value associated with imaginary sources.

## 5. Conclusion

This paper proposed a new sound source localization method using a small number of microphones. For suppressing imaginary sound sources, we introduced the minimum filtering of the GCCs rather than summing all GCCs as the conventional SRP-PHAT method. The experiments were conducted in a real noisy environment. As the result, we demonstrated that the proposed method can localize multiple sound sources more accurately in full azimuth than the conventional methods. In future work, we will extend the proposed method to 3-dimensions and with more microphones.

## Acknowledgments

This research was supported in part by Seoul R&BD (WR080951) and in part by a grant of Korea Health Technology R&D Project (A111189).

### References

[1] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," EURASIP J. Appl. Signal Process., vol.2006, pp.170–170, 2006.
[2] R. Schmidt, "Multiple emitter location and signal parameter-estimation," IEEE Trans. Antennas Propag., vol.34, no.3, pp.276–280, March 1986.
[3] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with arbitrarily arranged multiple sensors," J. Signal Process. Syst., pp.1–11, 2009.
[4] M. Cobos, J. Lopez, and S. Spors, "Analysis of room reverberation effects in source localization using small microphone arrays," Int. Symposium on Communications, Control and Signal Processing (ISCCSP), pp.1–4, Limassol, Cyprus, 2010.
[5] J.H. Dibiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. Thesis, AA (Brown University), Brown University, 2000.
[6] S. Rickard and Z. Yilmaz, "On the approximate W-disjoint orthogonality of speech," IEEE ICASSP, pp.529–532, Minneapolis, MN, USA, 2002.
[7] K. Cho, H. Okumura, T. Nishiura, and Y. Yamashita, "Multiple sound source localization based on inter-channel correlation using a distributed microphone system in a real environment," IEICE Trans. Inf. Syst., vol.E93-D, no.9, pp.2463–2471, Sept. 2010.
[8] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. Acoust. Speech Signal Process., vol.24, no.4, pp.320–327, 1976.
[9] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," IEEE Trans. Speech Audio Process., vol.5, no.3, pp.288–292, 1997.
[10] J. Beh, T. Lee, D. Han, and H. Ko, "Sound source separation by using matched beamforming and time-frequency masking," IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.458–463, 2010.
[11] Y. Sasaki, S. Kagami, and H. Mizoguchi, "Multiple sound source mapping for a mobile robot by self-motion triangulation," IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.380–385, 2006.
[12] M. Cobos, A. Marti, and J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," IEEE Signal Process. Lett., vol.18, no.1, pp.71–74, Jan. 2011.
[13] Z. Cha, D. Florencio, D. Ba, and Z. Zhengyou, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," IEEE Trans. Multimed., vol.10, no.3, pp.538–548, 2008.
[14] A. Marti, M. Cobos, and J. Lopez, "Real time speaker localization and detection system for camera steering in multiparticipant videoconferencing environments," IEEE ICASSP, pp.2592–2595, Prague, Cech, 2011.
[15] H. Silverman, Y. Ying, J. Sachar, and W. Patterson, "Performance of real-time source-location estimators for a large-aperture microphone array," IEEE Trans. Speech Audio Process., vol.13, no.4, pp.593–606, 2005.
[16] Y. Tamai, S. Kagami, Y. Amemiya, Y. Sasaki, H. Mizoguchi, and T. Takano, "Circular microphone array for robot's audition," Proc. IEEE Sensors, pp.565–570, 2004.
[17] K. Hayashida, M. Morise, and T. Nishiura, "Near field sound source localization based on cross-power spectrum phase analysis with multiple microphones," Interspeech, pp.2758–2761, Makuhari, Chiba, Japan, 2010.
[18] P. Aarabi, "The fusion of distributed microphone arrays for sound localization," EURASIP J. Appl. Signal Process., vol.2003, pp.338–347, 2003.
[19] J. Beh, T. Lee, I. Lee, H. Kim, S. Ahn, and H. Ko, "Combining acoustic echo cancellation and adaptive beamforming for achieving robust speech interface in mobile robot," IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.1693–1698, 2008.
[20] Y. Hioka and N. Hamada, "DOA estimation of speech signal using microphones located at vertices of equilateral triangle," IEICE Trans. Fundamentals, vol.E87-A, no.3, pp.559–566, March 2004.

**Suwon Shon** received the B.S. degree in Electrical Engineering from Korea University, Seoul, Korea, in 2010. Since 2010, he has been a Ph.D. course in Korea University. His research interests are in human-robot interaction, pattern recognition and multi-channel acoustic processing.

**David K. Han** received B.S. degree from Carnegie Mellon University, and MSE and Ph.D. from the Johns Hopkins University. After years of serving as scientist at NSWC and ONR, he joined the University of Maryland in College Park in 2005 as a visiting associate professor and the deputy director of the Center for Energetic Concepts Development (CECD). From 2007 to 2009, he was the Distinguished IWS chair professor in the Systems Engineering Department of the United States Naval Academy in Annapolis, MD. In Jan. 2009, Dr. Han returned to the ONR as a program officer in the Coastal and Geoscience (CG) team.

**Jounghoon Beh** received the B.S. degree M.S. degree, and Ph.D. Degree in Electronics and Computer Engineering from Korea University, Seoul, Korea, in 2001, 2003, and 2008, respectively. From 2008–2009, he has been a post-doctor in Korea University. Since 2010, he has been a post-doctor in Univ. of Maryland Institute for Advanced Computer Studies. His research interests are in human-robot interaction, hand gesture recognition, speech recognition, and multi-channel speech processing.

**Hanseok Ko** received B.S. degree from Carnegie Mellon University, in 1982, M.S. degree from the Johns Hopkins University, in 1988, and Ph.D. degree from the CUA, in 1992, all in electrical engineering. At the onset of his career, he was with the WOL, Maryland, where his work involved signal and image processing. In March of 1995, he joined the faculty of the Department of Electronics and Computer Engineering at Korea University, where he is currently Professor. His professional interests include speech/image signal processing for pattern recognition, multi-modal analysis, and intelligent data fusion.