Abnormal Acoustic Event Localization based on Selective Frequency Bin in High Noise Environment for Audio Surveillance

Suwon Shon
School of Electrical Engineering, Korea University
Seoul, Korea
swshon@ispl.korea.ac.kr

David K. Han
Ocean Engineering and Marine Systems Team
Office of Naval Research
Arlington, VA, USA
ctmkhan@gmail.com

Hanseok Ko
School of Electrical Engineering, Korea University
Seoul, Korea
hsko@korea.ac.kr

Abstract

In this paper, a method for source localization for surveillance system is presented. In particular, we propose an algorithm for abnormal acoustic event localization based on a novel approach of relevant frequency bin selections by statistical analyses. By means of selective frequency bin, it becomes possible to localize the event more accurately in high noise environment with low computational complexity. The effectiveness is verified through the experimental results in varied noise environments with different levels of Signal to Noise Ratio (SNR).

1. Introduction

Source localization has been an active research area and its techniques have been applied in many fields and applications [1]. In recent years, it became an important topic in surveillance systems related to acoustic processing such as sudden noise detection and localization in car, gunshot or scream detection [2], [3]. For practical applications, however, the technique has to be able to handle wideband signals, and yet is robust to highly noisy environment while the computational complexity is low.

Source localization is categorized into two classes. One of them is based on Direction Of Arrival (DOA) estimation. For high-resolution DOA estimation, many algorithms such as Maximum Likelihood (ML) method, ESPRIT have been developed [4], [5]. These methods have shown to be effective in simultaneous localization of multiple sources. However, they require high computational load and are not applicable to wideband sources [6], [7]. Other approaches such as the Multiple SIgnal Classification (MUSIC) method [8] developed for handling wideband sources, have limitations such as the number of sources has to be strictly less than the number of sensors [9]. They also exhibit difficulties in low SNR situation.

The other class is the Time Delay Estimation (TDE). For TDE, Generalized Cross Correlation (GCC) and Steered Response Power - PHAse Transform (SRP-PHAT) algorithms are developed [10–13]. Their time delay of arrival estimation is generally in low-resolution, but the method is robust against noise and reverberation compared to the DOA estimation and has low computational complexity. For these reasons, they are generally easier to implement to actual acoustic processing applications. More importantly, since it is suitable to wideband sources, a number of variations of the approach have been explored [14–18].

Denda et al. proposed weighted - Cross-power Spectrum Phase (CSP) coefficients based on an average speech spectrum [19]. Ichikawa et al. proposed a harmonic structure based weighting method for robust speech source localization [20]. Both of these methods have shown success in source localization for speech sources. Since they are designed for speech sources, however, they are not suitable for non-speech acoustic events such as breaking glass or alarm sounds.

Nakano proposed a Cross-Power Spectrum (CPS) method for improved GCC-PHAT by comparing a CPS of each frame with threshold and sub-band selection [21]. However, the CPS method required additional computational load due to the CPS computation. The method involves a threshold determined by three components: average noise level of input signal, highest CPS peak value in the input signal frame, and empirical gain factor G. With an initially determined average noise level, the performance may not be guaranteed because the CPS method not only selects the signal dominant sub-band, but also selects noise dominant sub-band when the SNR is lowered. This potential sensitivity to noise level hasn't been fully explored in this work. For robust performance, analyses of estimating average noise level of input frame and optimizing G value have to be explored.

To address these issues of wideband input signal of non-speech acoustic events in noisy environment, we propose a robust abnormal acoustic event localization algorithm using Selective Frequency Bin (SFB) method in wideband signal. The frequency bin group of each abnormal event class is determined by statistical analyzing of database. Using the selected frequency bin group, the proposed method achieves low computation complexity and remains robust in high noise environment by considering frequency bins that are closely associated with

978-1-4799-0703-8/13/$31.00 ©2013 IEEE

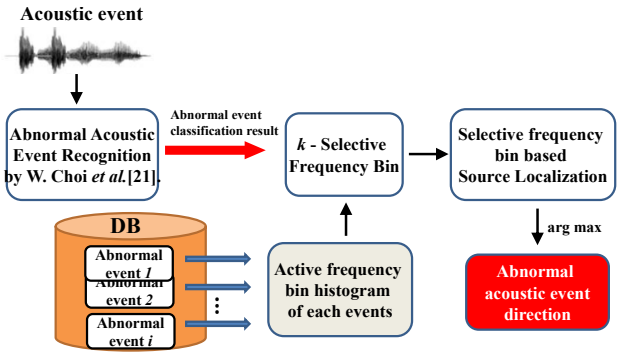


Figure 1: Block diagram of the proposed abnormal acoustic event localization system

abnormal acoustic events. For validating low computational requirement and robustness in noise, we selected six abnormal events representative in surveillance situations. Performance of the proposed method was evaluated by localizing 9 abnormal events with corresponding SNR and reverberation variations.

2. Abnormal Acoustic Event Localization

An overview of the proposed approach is sketched in Figure 1. First, an abnormal acoustic event recognition step is needed. A hierarchical structure based abnormal acoustic event recognition approach, developed by Choi, is adopted in this study [22]. After recognizing the acoustic event into one of the pre-defined abnormal events, *k* - SFB was determined based on the active frequency bin histogram corresponding to the abnormal event. Finally, source localization is performed based on the SFB, which eventually leads to the direction of an abnormal acoustic event.

For source localization, we use the TDE based algorithm. As we pointed out in the previous section, DOA estimation based algorithms generally has high angular resolution. But its implementation is too complex for wideband signals. TDE, such as GCC, SRP-PHAT, has lower angular resolution than DOA estimation based algorithms, but it has the advantage of reduced computational load. Although higher directional resolution is desirable for multiple sources in most cases, for abnormal events, it is safe to assume that high directional resolution is not required.

2.1. Signal model

Consider the general situation in which there are M-channel microphone signals at a discrete time *t* that can be represented by

$$x_m(t) = h_m(t) * s(t-\tau_m) + n_m(t) \qquad (1)$$

where *s(t)* is a sound source, $h_m(t)$ is the impulse response from the sound source to the *m*-th microphone, *n(t)* is the addictive white Gaussian noise and τ_m is propagation delay from source to *m*-th microphone.

2.2. Conventional Source localization algorithm

A brief description of TDE based SRP-PHAT algorithm follows. The GCC-PHAT of *l*-th and *q*-th microphone signals is

$$R_{lq}(\tau_{lq}, n) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Psi_{lq}(\omega,n) X_l(\omega,n) X_q^{*}(\omega,n) e^{j\omega\tau_{lq}} d\omega \qquad (2)$$

where τ_{lq} is Time Delay Of Arrival(TDOA) of *l*-th microphone and *q*-th microphone, n is the frame index, X_l is the Short Time Fourier Transform (STFT) of the *l*-th microphone, and Ψ_{lq} denotes a weight function that performs well under realistic acoustical conditions. The PHAse Transform (PHAT) function is defined as $\Psi_{lq}(\omega,n) = 1/|X_l(\omega,n)X_q^{*}(\omega,n)|$. The SRP-PHAT algorithm output can be expressed using GCC-PHAT as:

$$P_n(\theta) = \sum_{l=1}^{M}\sum_{q=l+1}^{M}\int_{-\infty}^{\infty} \Psi_{lq}(\omega) X_l(\omega,n) X_q^{*}(\omega,n) e^{j\omega\tau_{lq,\theta}} d\omega \qquad (3)$$

where $\tau_{lq,\theta}$ is the TDOA between *l*-th and *q*-th microphones when the source is at θ degree. The θ that has maximum amplitude of the SRP-PHAT determines the location of the sound source.

2.3. Proposed algorithm - SFB based abnormal acoustic event localization

For SFB based abnormal acoustic event localization, we must group the frequency bins by analyzing each event class. For grouping, a histogram of frequency bins was obtained by the frequency bins with power in excess of the average power of corresponding frame in frequency domain. Then, the *k* largest frequency bins of histogram were grouped for the SFB based source localization.

Index (*i*)	Event (s_i)	Total Number	Total Duration (sec)
1	Scream(female)	243	572
2	Scream(male)	171	316
3	Barking	719	349
4	Breakage of Glass	127	112
5	Siren	786	610
6	Impact sound	229	194
7	Gunshot	139	31
8	Crying	258	114
9	Skidding sound	100	201

Table 1. Abnormal acoustic event database information

Recently, abnormal event detection was studied in [22–24]. In particular, the earlier study for abnormal acoustic events recognition by Choi [22] considered some representative acoustic events in scenes such as streets, subway platform and etc. We set the same 9 abnormal events and the database from Choi [22] as shown in Table 1 . Half was used for SFB and the other half was used for performance evaluation at Section 3. It is assumed that the abnormal event detection was done accurately prior to the source localization, thus the event class of input signals is known in advance.

The energy information of each frequency bin is used for finding the SFB of abnormal events. The STFT of an event source signal $s_i(t)$ can be expressed as $S_i(\omega,n)$. The active frequency information $S_i(\omega,n)$ corresponding each event class is

$$m_i(\omega,n)=\begin{cases}1, & S_i(\omega,n)>\dfrac{1}{2\pi}\displaystyle\int_0^{2\pi}S_i(\omega,n)d\omega \\ 0, & otherwise.\end{cases} \quad (4)$$

where ω denotes the frequency index and n is the frame index. The active frequency bin histogram of event class i is

$$a_i(\omega)=\sum_{n=1}^{N}m_i(\omega,n)\ . \quad (5)$$

Finally, we express SFB group into a vector by sorting top k bins from a_i

$$\boldsymbol{b}_{i,k}=\left\lfloor \underset{0<\omega<2\pi}{\arg\max_1}(a_i(\omega)),\ \underset{0<\omega<2\pi}{\arg\max_2}(a_i(\omega)),\ \ldots\ ,\ \underset{0<\omega<2\pi}{\arg\max_k}(a_i(\omega))\right\rfloor \quad (6)$$

where $argmax_n$ returns argument of the n-th largest value. Finally, the source localization algorithm based on SFB is

$$P_n(\theta)=\sum_{l=1}^{M}\sum_{q=l+1}^{M}\int_{\omega\in\boldsymbol{b}_{i,k}}\Psi_{lq}(\omega)X_l(\omega,n)X_q^{*}(\omega,n)e^{j\omega\tau_{lq,\theta}}d\omega\ . \quad (7)$$

3. Experiments

We conducted experiments in two distinct setups for performance evaluation. One is done in a controlled environment with additive white Gaussian noise with varying levels of SNR and reverberation time (RT_{60}) interval. The other one is in a real background noise environment.

The simulated room is 3.5x4.5x2.5m^3 and a Uniform Linear Array (ULA) consisting of 4 microphones with inter-spacing distance of 10cm to each other is located at the center of the room. In this environment, the database was created with an abnormal source at DOA of 30˚ (θ_s) at 1.2m from the ULA using half of abnormal acoustic event database as Table 1 that is not used for determining SFB at Section 2.3. The database was composed with a white Gaussian noise corresponding to SNR -5dB ~ 30dB in 16kHz sampling rate at 16-bit. For applying the room reverberation effect (RT_{60} = 100ms and 300ms), we generated the room impulse response from Lehmann and Johansson's image-source method [25]. The length of STFT is 1024 and the Hamming window is adopted.

The performance evaluated with two metrics, namely Root Mean Squared Error (RMSE) and Probability Of Success (POS). Each trial was considered a success if the estimated DOA is within the angular tolerance. In this paper, we set the angular tolerance of 30˚, so a trial was deemed successful when the estimated DOA is between the θ_s -30 and θ_s +30.

3.1. Under white Gaussian noise

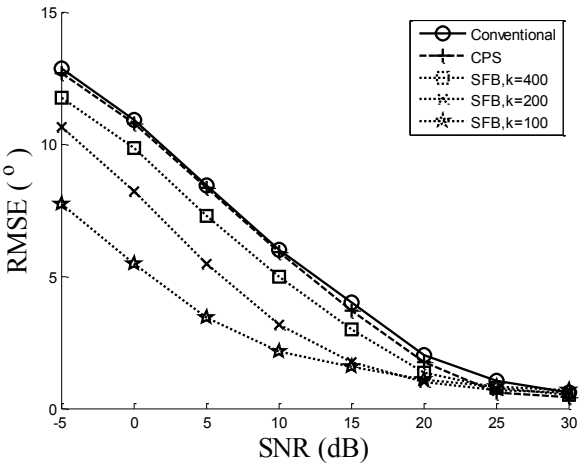
Figure 2 (a) and (b) shows RMSE and POS under additive white Gaussian noise with no reverberation. As shown from the figures, performance of the proposed approach (SFB) at k = 100 is much higher than that of the other algorithms in low SNR conditions. In high SNR, the performance was almost similar. However, in low SNR, the proposed approach only deals with the frequency bins wherein the abnormal sound source is dominant. The CPS method shows good performance in high SNR, but the approach shows performance degradation when SNR is lowered because sub-band selection based on CPS selects not only signal dominant sub-band but also selects noise concentrated sub-band.

Figure 2 (c) ~ (f) shows RMSE and POS evaluation result with 100ms and 300ms reverberation respectively. In both reverberation environments, POS of proposed algorithm has much higher probability than others. The RMSE evaluation also shows high accuracy of the proposed algorithm. If we select 100 frequency bins with a long reverberation such as at RT_{60} = 300ms, the RMSE is worse in high SNR. Nevertheless, it shows more accurate performance in low SNR.

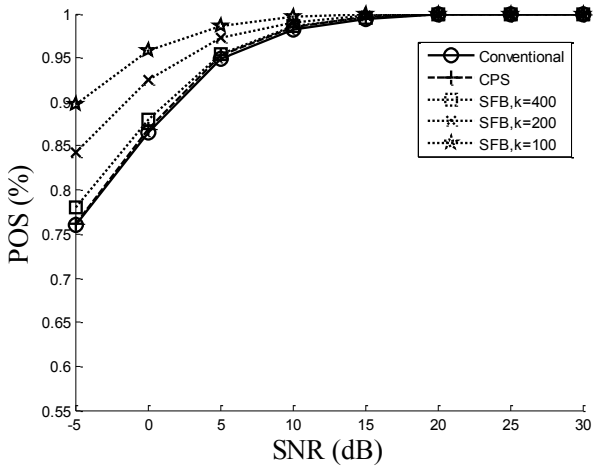
Average processing time per frame for each algorithm is compared with the conventional algorithm. The result in Table 2 shows the relative processing time between the proposed algorithms and the conventional with the processing time for conventional algorithm set as the baseline (100%). The SFB when k=200 and 100 has lower computational load than the CPS based algorithm while delivering high performance in high noise environment as shown Figure 2. Although it is not shown here, processing time for the SFB based algorithm remained mostly constant while that of the CPS based algorithm increased in high noise environment.

	CPS	SFB, k=400	SFB, k=200	SFB, k=100
Relative processing time	86.3%	90.6%	72.2%	62.5%

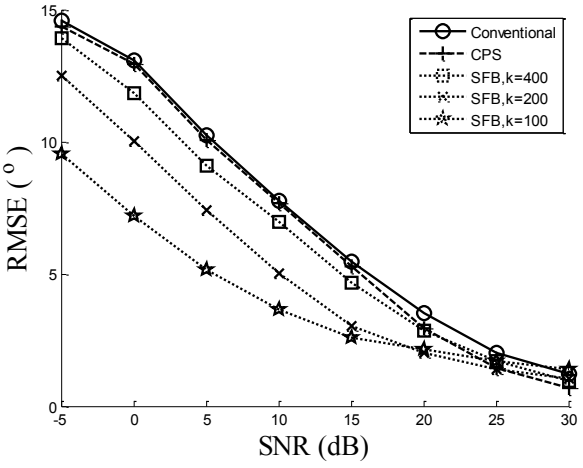
Table 2. Processing time of source localization



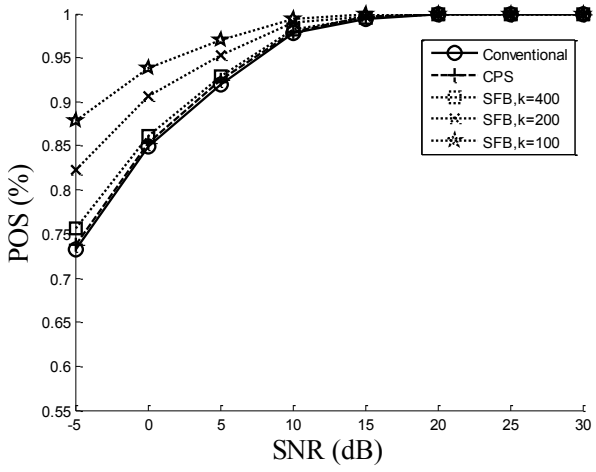
(a) no reverberation



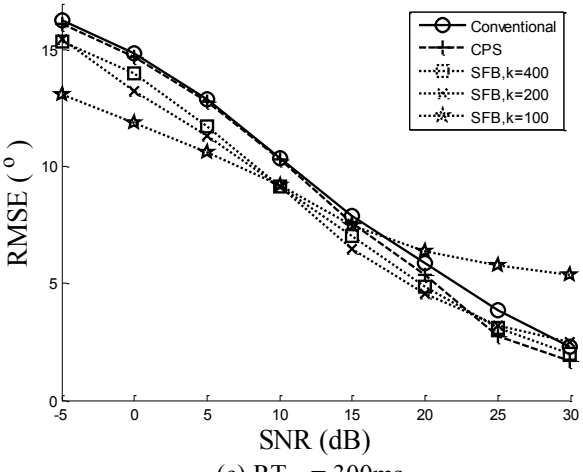
(b) no reverberation



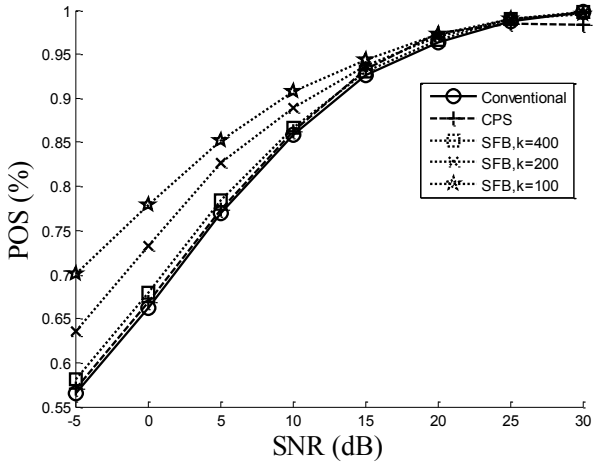
(c) RT_{60} = 100ms



(d) RT_{60} = 100ms



(e) RT_{60} = 300ms



(f) RT_{60} = 300ms

Figure 2: RMSE and POS of conventional, CPS and proposed algorithms corresponding reverberation under white Gaussian noise. (a) (b) : RMSE and POS with no reverberation, (c)(d) : RMSE and POS with RT_{60} =100ms, (e)(f) : RMSE and POS with RT_{60} =300ms

SNR	Background Noise	Conventional		CPS based		Proposed(SFB, k =200)	
		RMSE	POS	RMSE	POS	RMSE	POS
5dB	Silent	5.74°	86.8 %	4.34°	87.9 %	3.45°	91.2 %
	Street	9.11°	79.5 %	5.64°	83.9 %	5.49°	83.8 %
	Babble	7.85°	81.7 %	5.10°	83.3 %	4.26°	85.8 %
	Music	5.88°	86.6 %	4.92°	86.9 %	3.97°	88.6 %
	Subway	8.48°	83.9 %	5.20°	84.9 %	3.87°	86.7 %
	Water	10.77°	80.7 %	10.45°	79.7 %	8.75°	83.6 %
15dB	Silent	3.99°	93.4 %	3.69°	93.2 %	1.74°	96.7 %
	Street	6.99°	84.1 %	4.45°	85.6 %	4.02°	87.3 %
	Babble	6.13°	85.5 %	4.08°	85.8 %	2.69°	87.8 %
	Music	3.99°	90.1 %	3.74°	89.6 %	1.70°	95.8 %
	Subway	6.37°	86.5 %	4.28°	87.2 %	2.50°	90.4 %
	Water	8.20°	84.7 %	7.02°	85.0 %	5.53°	85.9 %
25dB	Silent	1.03°	97.2 %	0.61°	98.9 %	0.81°	99.9 %
	Street	4.88°	87.1 %	3.55°	88.0 %	2.82°	89.9 %
	Babble	4.76°	87.9 %	3.30°	89.6 %	1.67°	92.5 %
	Music	1.61°	97.1 %	1.42°	99.0 %	0.99°	99.8 %
	Subway	5.11°	90.3 %	2.61°	92.9 %	1.27°	96.7 %
	Water	5.98°	86.8 %	5.12°	87.3 %	4.13°	88.9 %

Table 3.RMSE and POS of conventional, CPS and proposed algorithms corresponding reverberation under white Gaussian noise under real noisy Environment with RT_{60} = 100ms

3.2. Under Real Background Noise

For validating the proposed algorithm, we conduct a feasibility test under a real background noise environment with the same database as [22] shown in Table 4. Using 6 different real background classes, we created multichannel abnormal acoustic event sources with remaining conditions identical to the ones in Section 3.1 including SNR at 5dB, 15dB and 25dB with RT_{60} = 100ms. Table 3 shows the average RMSE and POS in three SNR levels under 6 background noises. It shows that the proposed method performed the worst under water noise. This is due to the fact that water noise has flat frequency characteristics, e. g. uniformly distributed white noise, and it corrupts the entire frequency band. Considering the average of the 6 background noises, the RMSE shows similar performance compared to the one with white Gaussian noise. But it also shows about 8% decline in the POS compared to the case with white Gaussian noise.

Comparing the performance of the algorithms, the proposed algorithm, SFB with k = 200, seems to improve performance better than the other algorithms in both RMSE and POS. For RMSE, the proposed algorithm has 40% improvement compared to the conventional algorithm, and 25% improvement compared to the CPS algorithm. In POS, it shows 5% and 3% improvement compared to the conventional and CPS algorithms respectively. In particular, there is a significant performance improvement when SNR got worse. From these results, the proposed method seems quite feasible for practical applications in real environment.

Background Noise	Total Number	Total Duration (sec)
Silent place	750	2,556
Street	1,532	4,563
Babble place	2,990	7,968
Music	790	2,371
Subway	1,326	4,334
Water	1,741	5,221

Table 4. Background noise database information

4. Conclusions

Our proposed approach sets frequency bin groups that contain key acoustic features of the abnormal event. Then, we proposed SFB based source localization for acoustic event localization. Based on the set of experiments conducted, it demonstrated lower computational load while exhibiting more robust performance against noise than other conventional algorithms. Considering various performance evaluation, the SFB based source localization with k = 200 (about half of the frequency bins in a frame) is optimal solution for estimating DOA of abnormal sources in high noise environment. Its performance under real noise environment further validated that the abnormal acoustic event localization based on SFB is feasible in high noise environment for audio surveillance.

5. Acknowledgement

This research was supported by Seoul R&BD Program (WR080951).

References

[1] M. Brandstein and D. Ward, *Microphone Arrays : Signal Processing Techniques and Applications*. Springer, New York, 2001.

[2] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, 2007, pp. 21–26.

[3] S. Shon, E. Kim, J. Yoon, and H. Ko, "Sudden noise source localization system for intelligent automobile application with acoustic sensors," in *Consumer Electronics (ICCE), 2012 IEEE International Conference on*, 2012, pp. 233–234.

[4] I. Ziskind and M. Wax, "Maximum likelihood localization of multiple sources by alternating projection," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 36, no. 10, pp. 1553–1560, 1988.

[5] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989.

[6] Y.-S. Yoon, L. M. Kaplan, and J. H. McClellan, "TOPS: new DOA estimator for wideband signals," *IEEE Trans. on Signal Processing Letters*, vol. 54, no. 6, pp. 1977–1989, 2006.

[7] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Source counting in real-time sound source localization using a circular microphone array," in *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM),* , 2012, pp. 521–524.

[8] R. Schmidt, "Multiple Emitter Location and Signal Parameter-Estimation," *IEEE Trans. on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[9] Z. Wen-Jun and L. Xi-Lin, "High-Resolution Multiple Wideband and Nonstationary Source Localization With Unknown Number of Sources," *IEEE Trans. on Signal Processing*, vol. 58, no. 6, pp. 3125–3136, 2010.

[10] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[11] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1994, vol. ii, p. II/273–II/276 vol.2.

[12] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997, vol. 1, pp. 375–378 vol.1.

[13] J. H. Dibiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Brown university, 2000.

[14] H. Do, H. F. Silverman, and Y. Yu, "A Real-Time SRP-PHAT Source Location Implementation using Stochastic Region Contraction(SRC) on a Large-Aperture Microphone Array," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007, vol. 1, pp. I–121–I–124.

[15] 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 Processing Letters*, vol. 18, no. 1, pp. 71–74, 2011.

[16] J. P. Dmochowski, J. Benesty, and S. Affes, "A Generalized Steered Response Power Method for Computationally Viable Source Localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2510–2526, 2007.

[17] A. Johansson and S. Nordholm, "Robust acoustic direction of arrival estimation using Root-SRP-PHAT, a realtime implementation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005, vol. 4, p. iv/933–iv/936 Vol. 4.

[18] S. Shon, D. K. Han, J. Beh, and H. Ko, "Full Azimuth Multiple Sound Source Localization wigh 3-Channel Microphone Array," *IECE Trans. on Fundamentals*, vol. E95-A, no. 4, pp. 745–750, 2012.

[19] Y. Denda, T. Nishiura, and Y. Yamashita, "Robust Talker Direction Estimation Based on Weighted CSP Analysis andMaximum Likelihood Estimation," *IEICE Trans. on Information and Systems*, vol. E89-D, no. 3, pp. 1050–1057, 2006.

[20] O. Ichikawa, T. Fukuda, and M. Nishimura, "DOA Estimation with Local-Peak-Weighted CSP," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, pp. 1–10, 2010.

[21] A. Y. Nakano, S. Nakagawa, and K. Yamamoto, "Automatic estimation of position and orientation of an acoustic source by a microphone array network," *The Journal of the Acoustical Society of America*, vol. 126, p. 3084, 2009.

[22] W. Choi, J. Rho, D. K. Han, and H. Ko, "Selective Background Adaptation Based Abnormal Acoustic Event Recognition for Audio Surveillance," *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, pp. 118–123, Sep. 2012.

[23] D. Conte, P. Foggia, G. Percannella, A. Saggese, and M. Vento, "An Ensemble of Rejecting Classifiers for Anomaly Detection of Audio Events," *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, pp. 76–81, Sep. 2012.

[24] S. Lecomte, R. Lengellé, C. Richard, F. Capman, and B. Ravera, "Abnormal events detection using unsupervised One-Class SVM-Application to audio surveillance and evaluation," *... -Based Surveillance ...*, pp. 124–129, 2011.

[25] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 269–277, 2008.