TDOA Estimation Based on Phase-Voting Cross Correlation and Circular Standard Deviation

Masanori Kato, Yuzo Senda and Reishi Kondo
Data Science Research Laboratories, NEC Corporation
Shimonumabe 1753, Nakahara-ku, Kawasaki, Kanagawa, Japan

***Abstract*—This paper proposes a new TDOA estimation based on phase-voting cross correlation and circular standard deviation. Based on phase delay and kernel function, the proposed method generates a probability density function (PDF) of TDOA for each frequency bin. TDOA estimate is determined by voting the PDFs generated for all frequency bins. Peak positions of the bin-wise PDFs for the target signal are concentrated only at the target time difference because peak positions for the noise totally differ among bins and periodicity of peaks depends on frequency. Therefore, by voting the PDFs for all frequency bins, the peak position for the target can be easily identified. The kernel width of PDF is determined by circular standard deviation of cross spectral phase for each frequency bin. This width control enhances peaks of PDFs for high SNR frequency bins since phases for high SNR bins are more stable than those for low ones. Evaluation with ship and drone sounds shows that the RMSE of TDOA estimation by the proposed method reaches 0.37 times that by GCC-PHAT.**

I. Introduction

Acoustic environment understanding for detection of accidents or crimes is an important technology for establishing a safe and secure city [1]. Acoustic surveillance, which reports an incident and its time and place to an administrator or a supervisor based on analysis of observed acoustic signals, is one of applications of acoustic environment understanding. Acoustic surveillance is mainly supported by two functions: acoustic event detection and sound direction-of-arrival (DOA) estimation. Acoustic event detection recognizes and classifies a physical phenomenon or sound source which causes the observed event sound. Some of recent methods are found in [2]–[4]. DOA estimation is also important for acoustic surveillance. Its estimate is used for localization of the sound source, angle view control of surveillance cameras, and enhancement of the target sound for acoustic event detection.

Several DOA estimation methods have been proposed mainly for speaker direction estimation used in digital video cameras, teleconference systems and interactive robots. Among methods using many microphones, steered response power (SRP) localization [5] which finds a direction maximizing output power of steered beamformer, and subspace based methods [6], e.g., the MUSIC [7] algorithm, are common. For better performance, they usually require an array of many microphones, which causes limitation of surveillance place due to the large array.

For surveillance application, time difference of arrival (TDOA) estimation based on generalized cross correlation (GCC) [8] which basically uses only two microphones is widely used for DOA estimation [9]–[12]. To obtain the best TDOA estimate, simple exhaustive search for selecting pairs of microphone from the array is proposed [9]. Other conventional methods [10]–[12] improve estimation accuracy by selecting or weighting target signal frequency components estimated with cross spectral power. In [10], minimum variance distortionless response (MVDR) is used to estimate the target cross spectrum for GCC. In SNR estimation used in [11] for TDOA estimation, time-averaged input signal is adopted as target signal estimate. This approach is effective under an assumption that the target signal is stationary and has large power as compared with noise. SNR-based target sound onset detection in [12] extracts stationary components in the input power spectrum as estimated noise and uses them for detection.

These conventional methods work well under relatively high SNR conditions since they discriminate the target signal and noise by using power spectrum. High SNR conditions, however, do not always hold for actual surveillance environments. In addition, target sound power becomes smaller if target sound source is located far from the surveillance microphone position. Therefore, estimation accuracy improvement in lower SNR conditions is important for the surveillance application. However, it is difficult for conventional methods to archive sufficient estimation accuracy in low SNR conditions.

In this paper, we propose a new TDOA estimation method based on voting of bin-wise probability density function (PDF) of TDOA. The bin-wise PDF is derived from cross spectral phase and kernel function which give the center and the width of the PDF. The PDF of TDOA in total is given by the sum of all voted bin-wise PDFs. In the next section, the conventional GCC-PHAT is reviewed with its drawback. Section 3 explains the proposed method. Finally, in Section 4, evaluation results show superiority of the proposed method.

II. Conventional method

GCC with PHAse Transform (GCC-PHAT) [8] has been widely used for TDOA estimation and it is a basis for many acoustic source localization methods. Figure 1 shows a block diagram of GCC-PHAT. TDOA estimate $\hat{\tau}$ of two input signals is given by

$$\hat{\tau} = \arg\max_{\tau} r_{12}(\tau), \tag{1}$$

where

$$r_{12}(\tau) = \frac{1}{K}\sum_{k=0}^{K-1} \frac{S_{12}(k)}{|S_{12}(k)|} e^{j\frac{2\pi\tau k}{K}}, \tag{2}$$

ISBN 978-0-9928626-7-1 © EURASIP 2017

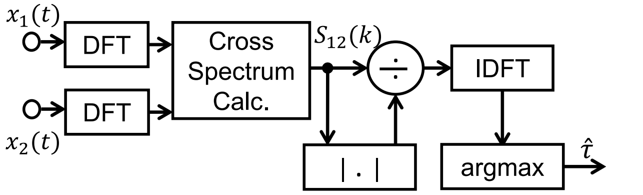

Fig. 1. Block diagram of GCC-PHAT

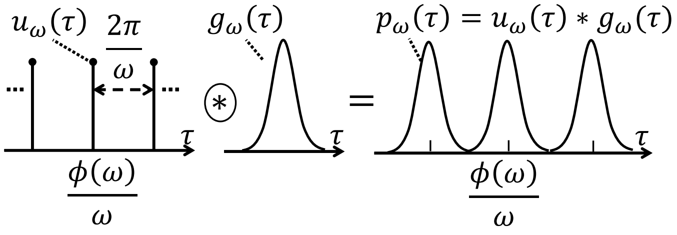
$u_\omega(\tau)$
$\frac{2\pi}{\omega}$
$g_\omega(\tau)$
$p_\omega(\tau) = u_\omega(\tau) * g_\omega(\tau)$
τ
$\frac{\phi(\omega)}{\omega}$

Fig. 2. Bin-wise PDF for frequency ω

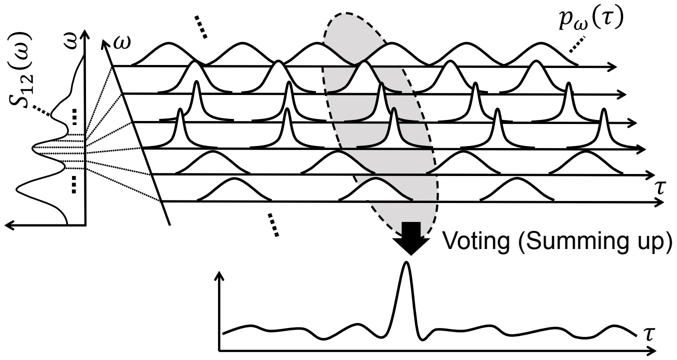

Fig. 3. Voting to aggregate bin-wise PDFs

k is a frequency bin index, K is the DFT length, and $S_{12}(k)$ denotes the cross spectrum of input signals. The main feature of GCC-PHAT is normalization applied to the cross spectrum. TDOA estimate is obtained as the peak position in the IDFT of the normalized cross spectrum. TDOA estimation by GCC-PHAT works with low computational cost, archives good estimation accuracy under low noise environments, even with high reverberation [13], [14].

However, estimation accuracy degrades when SNR becomes lower. In the case, the number of noise-dominant frequency bins increases, which results in injection of large noise into the summation as shown in equation (2). To cope with this problem, high SNR band selection based on cross spectral power is proposed as an enhancement in [10]–[12]. But it is not easy because of needs for accurate noise spectrum estimation or other strict assumption regarding the target signal or noise.

III. Proposed method

Unlike GCC's signal processing approach, a statistical approach is employed in the proposed method. In the approach, the cross spectral phase in each frequency bin is considered as a stochastic variable following a certain circular distribution. The center of the distribution is given by the expected value of the cross spectral phase. The width of the distribution depends on the reliability of the bin, in other words, reliable bins have a sharp distribution. In the proposed method, the distribution in the phase domain is translated into a bin-wise probability density function (PDF) in the time difference domain. The PDF of TDOA in total is given by the sum of all voted bin-wise PDFs like kernel density estimation. As the proposed method is based on phase and voting and its resulting PDF is a similar to a cross correlation function, it is named Phase-Voting Cross Correlation.

A. Bin-wise PDF

A bin-wise PDF is derived as the convolution of a frequency-dependent periodic function and a kernel function which give the center and the width of a circular distribution, respectively. Use of the periodic function makes phase unwrapping unnecessary. Figure 2 shows a bin-wise PDF at an angular frequency ω. In the figure, $\phi(\omega)$ is a cross spectral phase of the input signals, $u_\omega(\tau)$ is a periodic function generated with $\phi(\omega)$, $g_\omega(\tau)$ is a kernel function, and $p_\omega(\tau)$ is a PDF which is the convolution of $u_\omega(\tau)$ and $g_\omega(\tau)$. In the proposed method, Gaussian function is adopted for the kernel function.

B. Kernel width control based on circular standard deviation

The width of the kernel function is determined according to reliability of the frequency bin. For the bin whose reliability is high, the width is made narrower to enhance the peak of the bin-wise PDF. The proposed method employs circular standard deviation [15] as an index of reliability. For each frequency bin, circular standard deviation is calculated by using the cross spectra obtained in past frames. The width is then calculated based on the circular standard deviation, and applied to the kernel function generation. Kernel width control is effective when SNR varies across the bins. By considering time-series stability of cross spectral phase, high SNR bins can be identified.

Circular standard deviation is known as one of the common measures of circular spread, and a key parameter of wrapped normal distribution. As a bin-wise PDF is derived as the convolution of a periodic function and the Gaussian function, a cycle of the PDF is equivalent to the wrapped normal distribution around the circumference of a circle. Therefore, circular standard deviation can be used directly as the width of the kernel function if Gaussian kernel is used.

C. Voting of bin-wise PDF

TDOA estimate is determined by voting the bin-wise PDFs generated for all frequency bins as shown in Fig. 3. Peak positions of the bin-wise PDFs for the target signal are concentrated only at the target time difference because peak positions for the noise totally differ among bins and periodicity of peaks depends on frequency as shown in Fig. 2. Besides, the peak of a bin-wise PDF for the target signal is sharper than that for the noise thanks to the kernel width control. Therefore, by voting the bin-wise PDFs, the peak position for the target can be easily identified.

ISBN 978-0-9928626-7-1 © EURASIP 2017

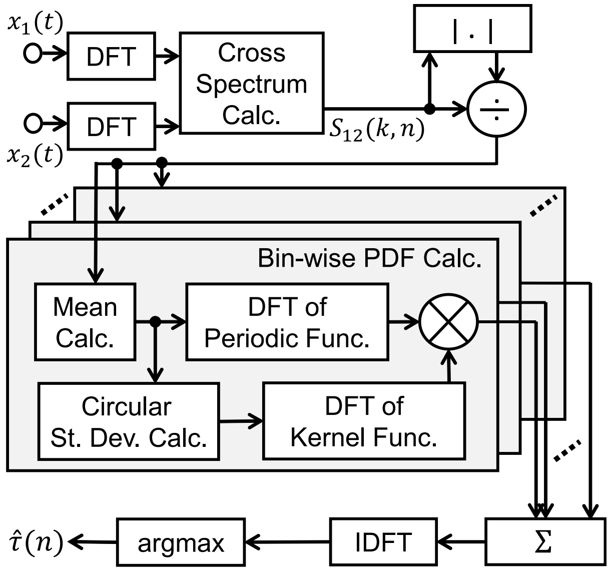

Fig. 4. Block diagram of proposed method

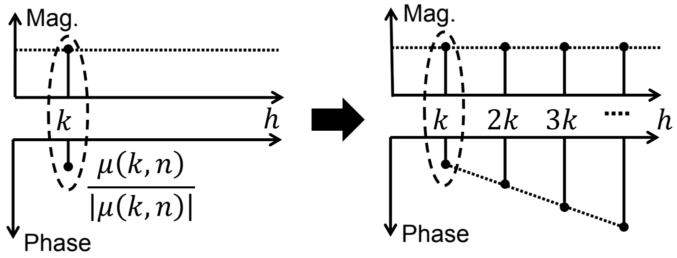

Fig. 5. DFT of periodic function

D. Detailed procedure

Figure 4 shows a block diagram of the proposed method. Compared to Fig. 1, blocks until normalization of cross spectrum are the same as the conventional method. To avoid convolution in the time domain, bin-wise PDF calculation and voting are conducted in the frequency domain.

The input signals are first divided into frames, then cross spectrum $S_{12}(k,n)$ is calculated for each frame, where n is a frame number. In bin-wise PDF calculation, mean $\mu(k,n)$ and circular standard deviation $\sigma(k,n)$ are given by

$$\mu(k,n) = \frac{1}{L}\sum_{l=0}^{L-1}\frac{S_{12}(k,n-l)}{|S_{12}(k,n-l)|}, \tag{3}$$

$$\sigma(k,n) = \sqrt{-2\log|\mu(k,n)|}. \tag{4}$$

The spectrum and time difference of the target are assumed stable during L frames. Based on phase linearization, the DFT of a periodic function, $U_k(h,n)$ where $h=0,1,\ldots,H-1$, is calculated with $\mu(k,n)$ as follows.

$$U_k(h,n) = \begin{cases} \left(\frac{\mu(k,n)}{|\mu(k,n)|}\right)^{h/k}, & h \bmod k = 0 \\ 0, & h \bmod k \neq 0 \end{cases} \tag{5}$$

Figure 5 illustrates the equation (5). $(\mu(k,n)/|\mu(k,n)|)^{h/k}$ is assigned for $U_k(h,n)$ in $h=k,2k,3k,\ldots$-th frequency bin to form linear phase. Zero is assigned for the other frequency bins. The DFT of a kernel function, $G_k(h,n)$, is given by

$$G_k(h,n) = \exp\left(-\frac{\sigma(k,n)^2}{2}\left(\frac{2\pi h}{H}\right)^2\right). \tag{6}$$

Kernel functions for various standard deviations can be calculated in advance to reduce computational complexity.

The DFT of a bin-wise PDF, $P_k(h,n)$, is given by

$$P_k(h,n) = U_k(h,n)G_k(h,n). \tag{7}$$

$P_k(h,n)$ is calculated for all frequency $k=0,1,\ldots,K-1$. DFT of total PDF, $P(h,n)$, is obtained by voting (summing-up) all bin-wise PDFs as follows.

$$P(h,n) = \sum_{k=0}^{K-1} P_k(h,n). \tag{8}$$

Finally, the maximizer of the IDFT of $P(h,n)$ is obtained as TDOA estimate.

In a special case that the kernel function is defined by

$$G_k(h,n) = \begin{cases} |\mu(k,n)|, & k = h \\ 0, & k \neq h \end{cases} \tag{9}$$

and $L=1$, then we have

$$P(h,n) = U_h(h,n)G_h(h,n) = \mu(h,n) = \frac{S_{12}(h,n)}{|S_{12}(h,n)|} \tag{10}$$

from Eqs. (3), (5), (7) and (8). Therefore, the proposed method can be considered as an extended version of GCC-PHAT.

IV. Evaluation

Evaluations were performed using two kinds of sounds recorded at 48 kHz in real environments, one of which is a small ship sailing in a bay, the other is a drone (small multicopter) flying in a countryside. Cross correlation function (CCF) of GCC-PHAT was compared as the conventional method to PDF of the proposed method, PVCC. The lengths of FFT and frame shift were 2048 and 1024, respectively. Estimated TDOA was updated by 0.5 seconds (23 frames) which was also used as averaging time for mean and circular standard deviation. Distances and directions shown in the following sections were obtained with measured GPS data.

A. Ship sound localization

Figure 6 illustrates the experiment setup for ship sound localization. During sound recording, a small ship sailed from port to sea. Recording time was 120 seconds. The direction of the ship was around 54 degree which is equivalent to 14 lag samples at 48 kHz sampling.

Figure 7 shows a spectrogram of the recorded sound. In the figure, frequency components of the ship sound appear in lower frequency surrounded by the dashed line and decay over time. As time goes on, the SNR of the ship sound decreases. Vertical line components in high frequency are noise sound generated by waves of the sea.

ISBN 978-0-9928626-7-1 © EURASIP 2017

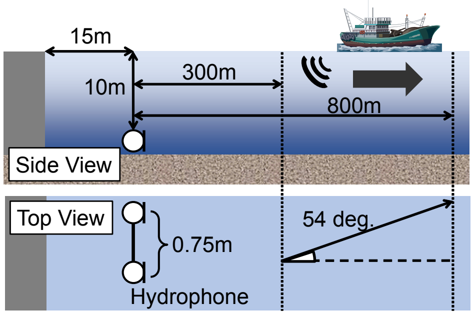

Fig. 6. Experiment setup for ship sound localization

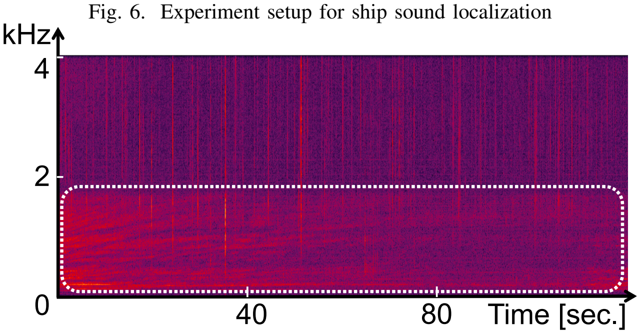

Fig. 7. Spectrogram of recorded sound (ship sound)

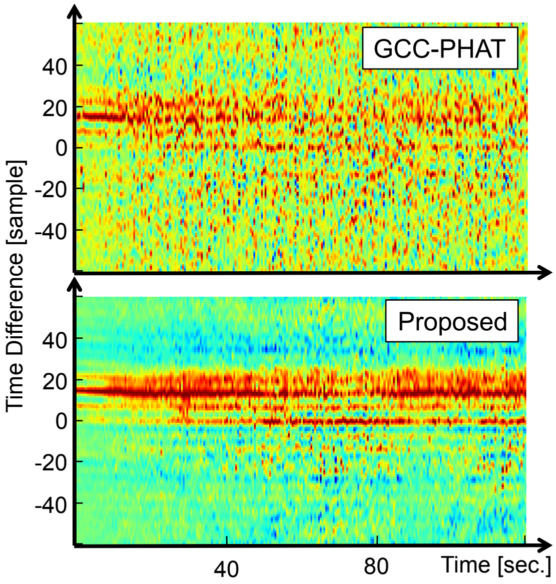

Fig. 8. CCF by GCC-PHAT and PDF by proposed method

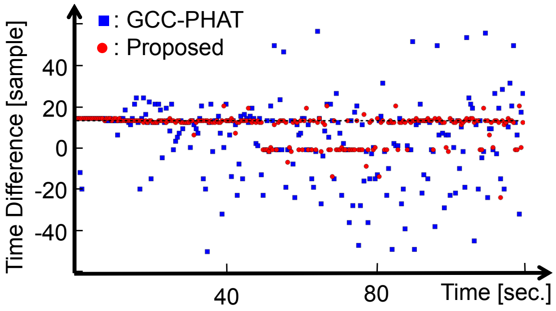

Fig. 9. TDOA estimated by GCC-PHAT and proposed method

Figures 8 and 9 show evaluation results. In Fig. 8, CCF and PDF are shown side by side. Comparing the two, PDF has a clear peak found around 14 lag samples for all the time while a clear peak of CCF appears only around the beginning. Figure 8 shows TDOA estimated by the conventional and the proposed methods. Root mean square errors (RMSE) of TDOA estimation by the conventional and the proposed methods are 0.42 msec. and 0.16 msec., respectively. These two values were calculated with the following equation.

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\tau_{true} - \hat{\tau}(n))^2}, \qquad (11)$$

where τ_{true} is the true TDOA, $\hat{\tau}(n)$ is the estimated TDOA at a frame n, N is the number of frames.

B. Drone sound localization

Figure 10 illustrates the experiment setup for drone sound localization. During recording, a drone hovered in the air with little movement. Recording time was 60 seconds. The direction of the drone was around -37 degree equivalent to -39 lag samples at 48 kHz sampling.

Figure 11 shows a spectrogram of the recorded sound. In the figure, frequency components of the drone sound mostly appear in lower frequency surrounded by the dashed line. Vertical line components in high frequency are of insect sounds. Figures 12 and 13 show evaluation results. In Fig. 12, CCF and PDF are shown side by side. Compared to CCF, PDF has clearer peak found around -39 samples for all the time. Figure 12 shows TDOA estimated by the conventional and the proposed methods. RMSEs of TDOA estimation by the conventional and the proposed methods are 0.53 msec. and 0.41 msec., respectively.

V. Conclusion

TDOA estimation based on phase-voting cross correlation and circular standard deviation has been proposed. Based on phase delay and kernel function, the proposed method generates a PDF of TDOA for each frequency bin. TDOA estimate is determined by voting the PDFs generated for all frequency bins. Peak positions of the bin-wise PDFs for the target signal are concentrated only at the target time difference because peak positions for the noise totally differ among bins and periodicity of peaks depends on frequency. Therefore, by voting the PDFs for all frequency bins, the peak position for the target can be easily identified. The kernel width of PDF is determined by circular standard deviation of cross spectral phase for each frequency bin. This width control enhances peaks of PDFs for high SNR frequency bins since phases for high SNR bins are more stable than those for low ones. Evaluation with ship and drone sounds has shown that the RMSE of TDOA estimation by the proposed method reaches 0.37 times that by GCC-PHAT.

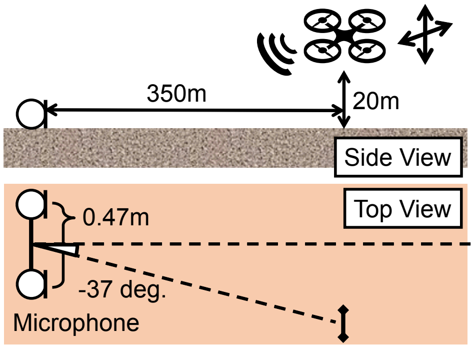

Fig. 10. Experiment setup for drone sound localization

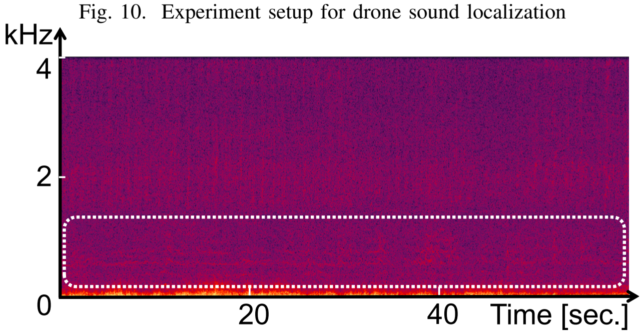

Fig. 11. Spectrogram of recorded sound (drone sound)

Acknowledgment

The authors would like to thank research members from Kanazawa Institute of Technology for their corporation.

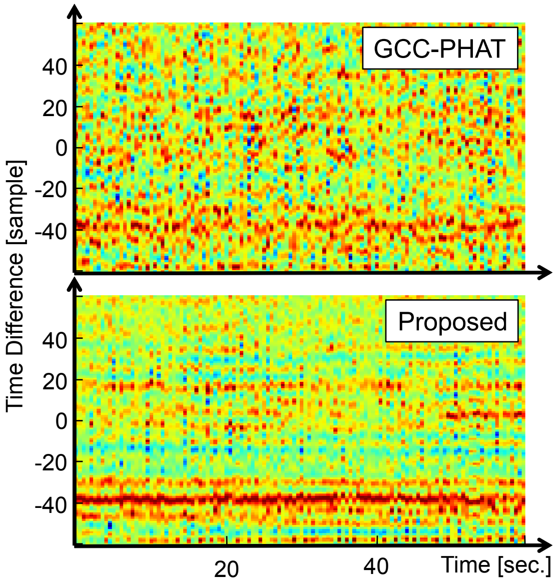

Fig. 12. CCF by GCC-PHAT and PDF by proposed method

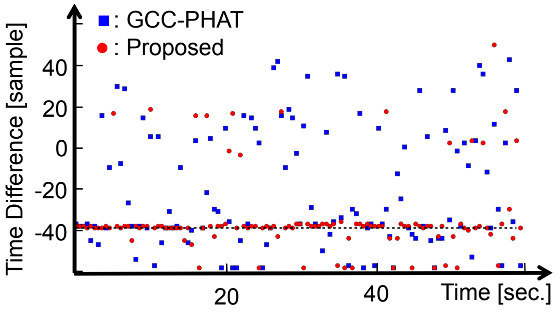

Fig. 13. TDOA estimated by GCC-PHAT and proposed method

References

[1] "NEC develops acoustic situation awareness technology that recognizes situations based on sound." http://www.nec.com/en/press/201611/global_20161128_03.html, Nov. 2016.

[2] C. V. Cotton and D. P. W. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 69–72, Oct. 2011.

[3] J. F. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste, and H. V. hamme, "An exemplar-based nmf approach to audio event detection," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–4, Oct. 2013.

[4] T. Komatsu, Y. Senda, and R. Kondo, "Acoustic event detection based on non-negative matrix factorization with mixtures of local dictionaries and activation aggregation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2259–2263, March 2016.

[5] J. Traa, D. Wingate, N. D. Stein, and P. Smaragdis, "Robust source localization and enhancement with a probabilistic steered response power model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 493–503, March 2016.

[6] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Doa estimation for multiple sparse sources with arbitrarily arranged multiple sensors," *Journal of Signal Processing Systems*.

[7] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, pp. 276–280, Mar 1986.

[8] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 320–327, August 1976.

[9] A. M. Borzino, J. A. Apolinario, and M. L. de Campos, "Robust doa estimation of heavily noisy gunshot signals," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 449–453, April 2015.

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

[11] C. W. Li and Y. W. Liu, "Posterior probabilistic modeling for inter-channel phase and time difference estimation in audio signals," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3206–3210, March 2016.

[12] P. Transfeld, U. Martens, H. Binder, T. Schypior, and T. Fingscheidt, "Acoustic event source localization for surveillance in reverberant environments supported by an event onset detection," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2629–2633, April 2015.

[13] C. Zhang, D. Florencio, and Z. Zhang, "Why does PHAT work well in low noise, reverberative environments ?," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2565–2568, March 2008.

[14] J. Velasco, M. J. Taghizadeh, A. Asaei, H. Bourlard, C. J. Martin-Arguedas, J. Macias-Guarasa, and D. Pizarro, "Novel GCC-PHAT model in diffuse sound field for microphone array pairwise distance based calibration," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2669–2673, April 2015.

[15] N. I. Fisher, *Statistical Analysis of Circular Data*. Cambridge University Press, 1993.