INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

A robust dual-microphone speech source localization algorithm for reverberant environments

Yanmeng Guo[1], *Xiaofei Wang*[1 2], *Chao Wu*[1], *Qiang Fu*[1], *Ning Ma*[3], *Guy J. Brown*[3]

[1]Institute of Acoustics, Chinese Academy of Sciences
[2] Center for Language and Speech Processing, Johns Hopkins University
[3]Department of Computer Science, University of Sheffield

guoyanmeng@mail.ioa.ac.cn, {wangxiaofei, wuchao, qfu}@hccl.ioa.ac.cn
{n.ma, g.j.brown}@sheffield.ac.uk

Abstract

Speech source localization (SSL) using a microphone array aims to estimate the direction-of-arrival (DOA) of the speech source. However, its performance often degrades rapidly in reverberant environments. In this paper, a novel dual-microphone SSL algorithm is proposed to address this problem. First, the time-frequency regions dominated by direct sound are extracted by tracking the envelopes of speech, reverberation and background noise. The time-difference-of-arrival (TDOA) is then estimated by considering only these reliable regions. Second, a bin-wise de-aliasing strategy is introduced to make better use of the DOA information carried at high frequencies, where the spatial resolution is higher and there is typically less corruption by diffuse noise. Our experiments show that when compared with other widely-used algorithms, the proposed algorithm produces more reliable performance in realistic reverberant environments.

Index Terms: Microphone array, Speech source localization, direction of arrival, reverberation.

1. Introduction

Speech source localization (SSL) aims to estimate the direction-of-arrival (DOA) of a speech source. It is important for voice capture [1] in many human-computer interaction applications, such as human-robot interaction, camera steering and intelligent monitoring.

Generally, the far-field assumption is applicable for a small-scale microphone array, so that the DOA can be estimated from the time difference of arrival (TDOA) or synchrony between the received signals. In methods based on a steered-beamformer [2], the peak output power is achieved once the signals are time-aligned. In algorithms derived from high-resolution spectral estimation [3], the spatial-spectral correlation matrix compensates for the time-delay difference between the received signals. TDOA can also be estimated based on inter-channel correlation [4], independent component analysis [5], zero-crossings [6], cross-power spectrum phase [7] and inter-channel phase difference (IPD) [8, 9].

Most SSL algorithms are reliable in free-field conditions, in which the received signal contains only the direct wave of the speech. However, in real application environments where room reflections occur, the captured signal inevitably contains both the direct sound and reverberation.

To achieve robustness in the presence of reverberation, the usual approach is to extract or emphasise the direct sound. To do so, some algorithms exploit the characteristics of the speech signal, such as its statistical independence from other sources [5], its harmonic structure [10], the excitation source of speech production [11, 12] and so on. Others attempt to cancel or eliminate the effect of the acoustic transfer function between the speaker and the microphones [2, 4, 8, 13, 14, 15] or utilize the consistency and continuity of the DOA in the frequency domain [16, 17] or time domain [18, 19].

High frequency parts of a signal are usually less corrupted by reverberation, because on average they have a higher absorption ratio. For example, phase transform (PHAT) weighting, which places equal importance on the phase of each frequency bin, has proven to be helpful in reverberant environments. However, high-frequency signals often cause spatial aliasing, which means that multiple wave cycles may be received at different microphones, and it turns the single-valued mapping between IPD and DOA into a multi-valued mapping.

Spatial aliasing can be avoided by discarding the high frequency signal or reducing the microphone spacing [20], but with consequent loss of localization resolution. Other methods utilize the redundancy contained in the received signal. For example, information from other frequency bands or time intervals [21, 22, 23] are reliable references or constraints. However, in applications with small microphone array scale, most references and constraints become inapplicable or unreliable.

In this paper, a dual-microphone based SSL algorithm is proposed to deal with reverberation for single speech scenario. The TDOA is estimated from time-frequency components which are dominated by the direct sound, and it is realized by an envelope tracking strategy for speech, reverberation and background noise. Then, a bin-wise de-aliasing method is proposed to remove the spatial aliasing, thus allowing high frequency bands to make a good contribution to the TDOA estimation.

2. Analysis of the problem

Consider an ideal anechoic environment containing a far-field speech source with spectrum $S(\omega)$. The received signal at microphone m ($m = 1, 2$) has the spectrum $X^{(m)}(\omega) = S(\omega)e^{-j\omega\tau_m}$, where τ_m is the time of propagation. Thus TDOA can be estimated correctly from the inter-channel phase difference, and the DOA is derived from $\sin\theta = \frac{c\delta}{d}$, where $\delta = \tau_1 - \tau_2$ is the TDOA, θ is the DOA, c is the speed of sound, and d is the inter-microphone distance.

In real environments, where reverberation and attenuation

http://dx.doi.org/10.21437/Interspeech.2016-1063

cannot be neglected, the received signal becomes

$$X(\omega) = a(\omega)S(\omega)e^{-j\omega\tau} + R(\omega) + D(\omega) \quad (1)$$

where $a(\omega)$ is the frequency-related attenuation, $R(\omega)$ is early reverberation, $D(\omega)$ is late reverberation, and the microphone index m is omitted. If we represent the reverberation as $R(\omega) + D(\omega) = N(\omega)e^{j\Phi_N(\omega)}$, then the phase of $X(\omega)$ is determined by $\omega\tau$ only if $|a(\omega)S(\omega)| >> |N(\omega)|$, which means that the estimated TDOA is close to its true value only if the time-frequency point is dominated by the direct sound. Therefore, the TDOA estimation is affected by the reverberation time T_{60}, which is the time required for reflections of a direct sound to decay 60 dB.

To illustrate the effect of reverberation, we calculate δ for each frequency bin and time frame to estimate the normalized histogram count of δ, which is depicted as $P(\delta)$. Fig. 1 shows $P(\delta)$ for speech and non-speech segments in environments with T_{60} = 300 ms and T_{60} = 600 ms respectively. The signal is 15 seconds long, containing 3 sentences and 4 intervals, and the DOA is $\theta_0 = 60°$. The distance between the two microphones is 0.085 m, so the true TDOA is $\delta_0 = 0.085 \cdot \sin 60°/c \approx 2.17 \times 10^{-4}$s. The sample rate is 16 kHz, and we use the Hann window, short-term Fourier transform (STFT) of 512 points and frame shift of 160 points. δ is estimated in frequency between 300 and 2000 Hz to avoid spatial aliasing.

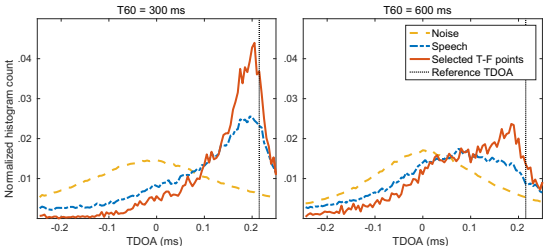

Figure 1: Histograms of the TDOA estimated from speech segments, intervals (noise) and selected time-frequency points in two reverberant environments. The reference TDOA $\tau_0 = 0.217$ ms is shown as vertical dotted lines.

As is shown in Fig. 1, $P(\delta)$ is highly affected by reverberation in speech segments, whereas it is relatively unaffected in noise segments. Higher reverberation reduces the differentiation between speech and noise, and cause higher bias and variance in TDOA estimation. Moreover, due to the mapping of $\sin\theta = \frac{c\delta}{d}$, higher bias or variance of δ will bring more serious bias for θ estimation.

Therefore, only reliable parts of received signal should be extracted for TDOA estimation. This is realized in two ways in this paper. First, the direct wave component is extracted by envelope tracking, which exploits the fact that direct sound arrives earlier than reflections, so $S(\omega)$ can dominate $X(\omega)$ on its rising edges, while the proportion of $R(\omega)$ and $D(\omega)$ usually increase later. Second, a sound wave with higher frequency usually decays faster than one with a lower frequency, thus it is more likely to be dominated by $S(\omega)$. To allow the use of high-frequency bands, an approach for eliminating spatial aliasing by appropriate selection of TDOA candidates is presented.

3. Algorithm description

The proposed algorithm can be summarised as follows. First, the time-frequency (T-F) points that carry the TDOA information are extracted. This is realized by the envelope tracking of the speech, early reverberation and background in their amplitude of cross-power spectrum. Secondly, a reliable TDOA estimator for high-frequency bands is described, and a bin-wise de-aliasing strategy is utilized to delete the aliased TDOA estimators. Finally, the DOA is estimated based on the distribution of the reliable TDOA estimators.

3.1. Envelope tracking

The signals received in two channels are transformed into the frequency domain via STFT, and depicted as $X_{m,l}(k)$, where m ($m = 1, 2$) is the channel index, l is the frame index, and k is the frequency bin. Then the amplitude of cross-power spectrum is calculated as $C_l(k) = |X_{1,l}(k)X^*_{2,l}(k)|$ and logarithmically compressed to $E_l(k) = \log_{10} C_l(k)$.

The envelopes are tracked in each frequency bin. Here we omit the index k, and denote $E_l(k)$ as E_l. Three envelopes are tracked based on E_l: direct speech S_l, early reverberation R_l, and ground noise G_l. Here the ground noise is the summation of all short-time stationary noises, including diffuse noise, circuit noise and the stationary noise from the environment.

S_l is actually the excitation of the whole system, so S_l is the major component in the rising edge of E_l, and it is updated according to equation (2). λ_S adjusts the decay time of the speech envelope, which is set as 0.1 s based on the typical length of syllables and the speech gaps. If the frame shift is x second, then $\lambda_S = x/0.1$.

$$\begin{aligned} &\text{if} \quad E_l \geq S_{l-1}, \quad S_l \leftarrow E_l \\ &\text{else} \quad S_l \leftarrow \lambda_S E_l + (1-\lambda_S)S_{l-1} \end{aligned} \quad (2)$$

R_l increases after S_l because of the delay in multi-path propagation, and it decreases more slowly. R_l is updated according to (3), where μ_R is to describe the delay of the reflections, and λ_R adjusts the decay time of reverberation. For the rising time of 0.02 s and decay time of 0.5 s, we set $\mu_R = x/0.02$ and $\lambda_R = x/0.5$.

$$\begin{aligned} &\text{if} \quad E_l \geq R_{l-1}, \quad R_l \leftarrow \mu_R E_l + (1-\mu_R)R_{l-1}, \\ &\text{else} \quad R_l \leftarrow \lambda_R E_l + (1-\lambda_R)R_{l-1} \end{aligned} \quad (3)$$

G_l increases slowly and decreases fast to catch the gaps between speech segments. G_l updates according to (4), where μ_G and λ_G are parameters that adjust the rise and decay times. Typically we set the rise time as 1 s and decay time as 0.1 s.

$$\begin{aligned} &\text{if} \quad E_l \geq G_{l-1}, \quad G_l \leftarrow \mu_G E_l + (1-\mu_G)G_{l-1} \\ &\text{else} \quad G_l \leftarrow \lambda_G E_l + (1-\lambda_G)G_{l-1} \end{aligned} \quad (4)$$

All the T-F points with $S_l < R_l$ or $S_l < G_l + \eta$ are eliminated, where η is a frequency-related threshold. The higher the frequency, the lower η becomes, because the energy of clean speech attenuates by 6 dB/octave.

The purpose of envelope tracking is to delete the trailing parts of speech. It can be regarded as a sieve to extract the time-varying components while ignoring the prolonged or stationary components. Therefore, S_l rises instantaneously in the rising edge to to extract direct speech, while the trailing part is controlled by the decay time of the three components. The final performance of the SSL algorithm is not very sensitive to the parameters chosen, especially those relating to the updating of R_l and G_l.

Fig. 2 is an example of envelope tracking, in which the speech is recorded in a room with size 6m × 5m × 3m and

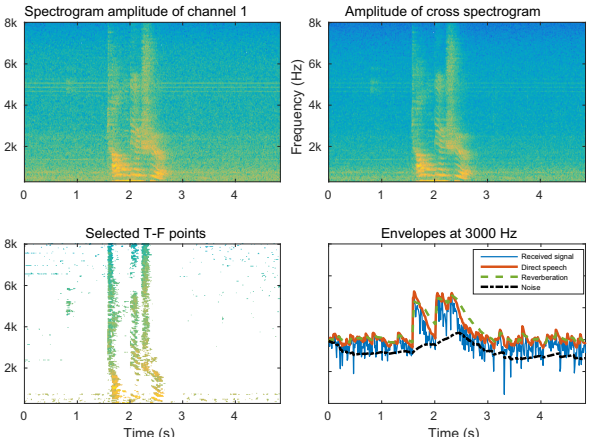

Figure 2: Illustration of envelope tracking.

$T_{60} \approx 600$ ms. The speaker is 3 m away from the microphones, and $d = 0.085$ m. The data is sampled at 16 kHz and analyzed with a frame shift of 0.01 s and STFT size of 512 points. The top right panel is the amplitude of cross-power spectrum of the received signal, in which the effect of diffuse noise is partly eliminated, especially at high frequencies. The bottom left panel shows the extracted region, where the T-F points dominated by direct sound are selected while most of the others are deleted. The bottom right panel displays the detail of envelope tracking for the frequency bin centered at 3000 Hz.

The effect of envelope tracking is also shown in Fig. 1 to compare the histogram of estimated δ in the extracted T-F parts and the ground truth (hand-labeled) speech segments. On the selected T-F parts, the peaks of the histograms are closer to the true value, and the peaks are higher and narrower. This means that the δ derived from the selected T-F points is closer to the true value, and this effect is more evident in the $T_{60} = 600$ ms environment. However, the peaks are still biased towards 0 in both environments, especially in the environment with a longer reverberation time. This is because most of the T-F points dominated by speech are still a little contaminated by the reverberation, which introduces a bias towards 0°. Therefore, the performance of SSL will not be reliable if it only relies on the information in the low frequency band.

3.2. TDOA de-aliasing

TDOA can be estimated for each T-F point based on IPD. Denote the phase of a T-F point on channel 1 and 2 as Φ_1 and Φ_2, where the frame and frequency index are omitted, then IPD is calculated as $\psi = \Phi_1 - \Phi_2$. So TDOA $\delta = \frac{\psi + 2n\pi}{2\pi f}$, where f is the frequency and n is an integer that satisfies

$$-\frac{d}{c} < \frac{\psi + 2n\pi}{2\pi f} < \frac{d}{c} \tag{5}$$

If $f > \frac{c}{2d}$, there may exist several values of n because of phase wrapping, but only one is correct. Therefore, TDOA de-aliasing is required in order to identify the correct n for $\delta = \frac{\psi + 2n\pi}{2\pi f}$.

According to (5), the distance between the two candidate δs is $\frac{\psi+2(n+1)\pi}{2\pi f} - \frac{\psi+2n\pi}{2\pi f} = 1/f$. This can be explained in two ways. First, if the possible δ range is limited to $1/f$, then the aliasing problem is avoided because only one n is possible. Second, the signal at frequency f has the best ability to differentiate δ in range with width $1/f$, because δ is just mapped to ψ of its full possible range $(0, 2\pi)$.

Therefore, lower frequencies are less affected by aliasing, but are not precise enough for TDOA estimation. On the other hand, higher frequency bands have better local precision, but the aliasing may be serious. To get good TDOA precision while keeping IPD un-aliased, a bin-wise de-aliasing algorithm is proposed here.

Assuming there is a single speech source, for a buffer with L frames, a TDOA distribution histogram $h_k(\delta)$ is first estimated based on all the selected reliable T-F points in the non-aliased frequency band, where k is the highest frequency bin of this band. Representing the frequency of bin k as f_k, then the range of δ in $h_k(\delta)$ can be denoted as $(\delta_k, \delta_k + \frac{1}{f_k})$. For the non-aliased frequency band, $\delta_k = -\frac{d}{c}$, and $\frac{1}{f_k}$ is equal to or a little higher than $\frac{2d}{c}$, according to the specific parameters.

In the higher bin $(k+1)$, the widest non-aliased range of δ is $(\delta_{k+1}, \delta_{k+1} + \frac{1}{f_{k+1}})$, where the start point δ_{k+1} should be determined to eliminate the range with ambiguity. The de-aliasing process in bin $(k+1)$ is deployed by searching the starting point in range of $[\delta_k, \delta_k + \frac{1}{f_k} - \frac{1}{f_{k+1}})$ based on the histogram $h_k(\delta)$, and the standard for the the chosen range is to have the highest summation of $h_k(\tau)$, as is shown in (6).

$$\delta_{k+1} = \arg\max_{\xi} \int_{\xi}^{\xi + \frac{1}{f_{k+1}}} h_k(\delta) \mathrm{d}\delta \tag{6}$$

For the L frames in the buffer, all the values of δ estimated from bin $(k+1)$ are wrapped to the range $(\delta_{k+1}, \delta_{k+1} + \frac{1}{f_{k+1}})$, by which the only one proper n is determined. Then the TDOA histogram is updated as $h_{k+1}(\tau)$ by introducing the T-F points on bin $(k+1)$. In the same way, the spatial aliasing for bin $(k+2)$ and higher frequency bands can also be eliminated, causing the TDOA histogram to become narrower and clearer.

The main idea in this de-aliasing strategy is to get a raw histogram of δ based on the un-aliased frequency band, which is utilized as the *a priori* distribution for the possible values of n in higher frequency bins. The de-aliased candidate of n is selected as the one with highest possibility, and a new histogram with a narrower range is formed by merging the new samples. Due to the bin-wise process, merging is reliable as long as the number of T-F points in the buffer is high enough.

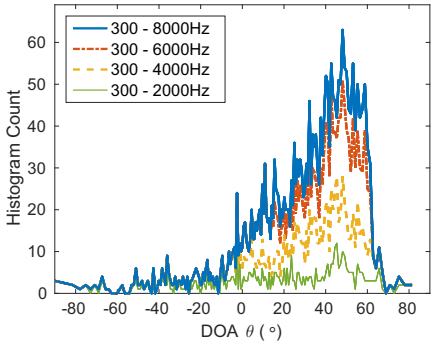

Figure 3: Illustration of the effect of TDOA de-aliasing.

Fig. 3 is an example of TDOA de-aliasing in a buffer of speech, where $d = 0.085$ m, the sample rate is 16 kHz, STFT size is 512, and $L = 25$. The TDOA is converted to DOA to show the effect more clearly. The non-aliased frequency band is 0-2 kHz, so the histogram of δ based on 300-2000 Hz is first calculated. The non-aliased histogram is low and flat, but the curve

becomes higher and clearer when progressively higher frequency bands are included. Finally, the DOA can be estimated as the one corresponding to the peak of the histogram.

4. Experiment and Analysis

4.1. Experimental setup

The performance of the SSL algorithm was tested on a corpus of signals recorded in a $6\,\mathrm{m} \times 5\,\mathrm{m} \times 3\,\mathrm{m}$ varechoic chamber. The T_{60} of the room could vary from 300 ms to 700 ms by adding or removing the sound absorbing panels on the walls. The speech data consisted of 64 clean Chinese sentences read by two men and two women, and the endpoints of speech were all hand-labeled. The speech was played by a loudspeaker 3 m away from the microphones with DOAs of 0°, 30°, 45° and 60°, respectively. Two omni-directional microphones with $d = 0.085\,\mathrm{m}$ were used to record the signals.

The received signals were sampled at 16 kHz, then Hann window weighted before applying a STFT of 512 points, with a frame shift of 0.01s. The frequency band below 300 Hz was discarded to avoid low-frequency interference. Then based on the frame shift, parameters of the proposed algorithms were set as below: $\lambda_S = 0.1$, $\mu_R = 0.5$, $\mu_G = 0.01$, $\lambda_G = 0.125$, and $L = 20$. Two values of λ_R were tested: 0.0333 and 0.0167, corresponding to the decay time of 300 ms and 600 ms.

4.2. Results and comparison

The proposed algorithm is compared with GCC, GCC-PHAT [4], SRP and SRP-PHAT [2] in terms of root-mean-square (RMS) error, as is shown in Table 1, where the rows **Proposed1** and **Proposed2** correspond to the proposed algorithm with λ_R=0.0333 and 0.0167 respectively.

Due to the frame-based processing in GCC and SRP, only the frames hand-labeled as speech are utilized in the RMS calculation. Moreover, a 7-frame post-processing is used to refine the localization result for each frame (i.e., the result of frame M is defined as the best result of frames $M-3$ to $M+3$).

As is shown in Table 1, all the algorithms have low bias when $\theta = 0^\circ$, regardless of the level of reverberation. However, the performance degrades when θ or reverberation becomes higher. The performance of GCC is affected by reverberation most seriously, followed by SRP, and the bias becomes higher when the DOA is higher. For both GCC and SRP, PHAT helps to reduce the bias in reverberant environments. The proposed algorithm shows the lowest bias when the DOA is not 0°, and the bias increases slowly with DOA and reverberation level.

4.3. Analysis of parameter values

The parameters in the proposed algorithm are set based on the property of speech signal and the propagation properties of sound waves, hence the performance of the algorithm should not be sensitive to the environment, so long as the parameters are within a reasonable range.

λ_R determines the decay time of the reverberation envelope, which can be set in the range of 200 ms to 1000 ms. As shown in Table 1, the change of reverberation envelope decay time from 300 ms to 600 ms only causes a small difference in the RMS error. Actually, a decay time close to the environment T_{60} will help to extract reliable T-F points in the trailing part of speech, but the final result is mainly determined by the rising edge because of the rapid decrease of the speech envelope.

The effect of different de-aliasing buffer length L was also tested. A longer buffer length is helpful for the final accuracy if θ remains stationary, because more reliable T-F points in low frequency bands will be involved to form the raw histogram of δ. However, if the buffer is too long, the algorithm will fail to estimate the instantaneous DOA of the speaker. Therefore, an appropriate buffer length should be selected according to the specific application to balance the accuracy and the tracking velocity, and the recommended range is between 200 and 300 ms.

Table 1: RMS error (in degrees) of the algorithms. Proposed1 and Proposed2 correspond to the proposed algorithm with reverberation envelope decay time of 300 ms and 600 ms.

T_{60}	Algorithm	DOA ($^\circ$)			
		0	30	45	60
300ms	GCC	3.07	16.40	23.81	40.11
	GCC-PHAT	2.25	7.17	10.93	16.40
	SRP	3.04	8.89	14.57	18.83
	SRP-PHAT	2.46	7.02	11.59	15.14
	Proposed1	**2.00**	**4.66**	**6.27**	**6.11**
	Proposed2	**2.11**	**4.79**	**6.99**	**6.13**
600ms	GCC	3.19	14.81	22.64	33.55
	GCC-PHAT	3.14	9.05	13.11	20.60
	SRP	2.28	11.22	21.91	30.81
	SRP-PHAT	2.84	8.67	16.41	23.86
	Proposed1	**1.73**	**7.34**	**11.52**	**13.39**
	Proposed2	**1.94**	**7.37**	**11.29**	**13.39**

5. Conclusions

A low-biased dual-microphone speech source localization algorithm is proposed in this paper. The T-F parts dominated by direct sound are extracted by an envelope tracking strategy motivated by the property of sound wave propagation. Then the aliased high frequency signal is fully utilized for TDOA estimation through a bin-wise de-aliasing process. Experiments show that the proposed algorithm has reliable performance in reverberant environments. Moreover, the algorithm is applicable to track a moving speech source if the buffer length is set to an appropriate value.

There are still some limitations of this algorithm. First, the de-aliasing process is based on the assumption that there exists only one speech source, and this condition is not always applicable in real applications. Second, the envelope tracking process is deployed in each frequency bin, and the correlation between different frequency bins could be further exploited.

In both respects, a strategy that groups correlated frequency bins will be helpful. Instead of the separate envelope tracking in each frequency bin, a contour tracking that involves several correlated frequency bins will be more practical in multi-source conditions to extract the direct sound, thus allowing the spatial de-aliasing strategy to be generalized to deal with multi-source conditions. This will be addressed in our future research.

6. Acknowledgements

This work is supported by the China Scholarship Council (No. 201504910055, 201604910007). Ma and Brown were supported by the EU FP7 project TWO!EARS under grant agreement 618075.

7. References

[1] I. Cohen and B. Berdugo, "Multichannel signal detection based on the transient beam-to-reference ratio," *IEEE Signal Processing Letters*, vol. 10, no. 9, pp. 259–262, 2003.

[2] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, 2000.

[3] C. T. Ishi, O. Chatot, H. Ishiguro, and N. Hagita, "Evaluation of a music-based real-time sound localization of multiple sound sources in real noisy environments," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009, pp. 2027–2032.

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

[5] A. Lombard, Y. Zheng, H. Buchner, and W. Kellermann, "TDOA estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1490–1503, 2011.

[6] Y.-I. Kim and R. M. Kil, "Estimation of interaural time differences based on zero-crossings in noisy multisource environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 734–743, 2007.

[7] M. Omologo and P. Svaizer, "Use of the crosspower spectrum phase in acoustic event location," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 5, no. 3, pp. 288–292, 1997.

[8] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.

[9] W. Zhang and B. D. Rao, "A two microphone-based aproach for source localization of multiple speech sources," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 1913–1928, 2010.

[10] M. S. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *Journel of Acoustical Society of America*, vol. 105, no. 5, pp. 2914–2919, 1999.

[11] V. C. Raykar, B. Yegnanarayana, S. R. M. Prasanna, and R. Duraiswami, "Speaker localization using excitation source information in speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 751–761, 2005.

[12] B. Yegnanarayana, S. M. Prasanna, R. Duraiswami, and D. Zotkin, "Processing of reverberant speech for time-delay estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, pp. 1110–1118, 2005.

[13] K. D. Donohue, J. Hannemann, and H. G. Dietz, "Performance of phase transform for detecting sound sources with microphone arrays in reverberant and noisy environments," *Signal Processing*, vol. 87, no. 7, pp. 1677–1691, 2007.

[14] R. Parisi, F. Camoes, M. Scarpiniti, and A. Uncini, "Cepstrum prefiltering for binaural source localization in reverberant environments," *IEEE Signal Processing Letters*, vol. 19, no. 2, pp. 99–102, 2012.

[15] T. Gustafsson, B. D. Rao, and M. Trivedi, "Source localization in reverberant environments: modeling and statistical analysis," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 791–803, 2003.

[16] Z. E. Chami, A. Guerin, A. Pham, and C. Servière, "A phase-based dual microphone method to count and locate audio sources in reverberant rooms," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA*, 2009, pp. 209–212.

[17] A. Cirillo, R. Parisi, and A. Uncini, "Sound mapping in reverberant rooms by a robust direct method," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2008, pp. 285–288.

[18] J. Benesty, J. Chen, and Y. Huang, "Time-delay estimation via linear interpolation and cross correlation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 509–519, 2004.

[19] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

[20] V. V.Reddy, B. P. Ng, Y. Zhang, and A. W. H. Khong, "DOA estimation of wideband sources without estimating the number of sources," *Signal Processing*, vol. 92, no. 4, pp. 1032–1043, 2012.

[21] L. Wang, H. Ding, and F. Yin, "A region-growing permutation alignment approach in frequency-domain blind source separation of speech mixtures," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 549–557, 2011.

[22] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Solving the permutation problem of frequency-domain bss when spatial aliasing occurs with wide sensor spacing," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2006, pp. V77–V80.

[23] R. Shimoyama and K. Yamazaki, "Computational acoustic vision by solving phase ambiguity confusion," *Acoustical Science and Technology of Japan*, vol. 30, pp. 199–208, 2009.