Direction of Arrival Estimation Based on Reverberation Weighting and Noise Error Estimator

Cheng Pang[2], *Jie Zhang*[2], *Hong Liu*[1,2]

[1]Key Laboratory of Machine Perception (Ministry of Education),
[2]Engineering Lab on Intelligent Perception for Internet of Things (ELIP),
Shenzhen Graduate School, Peking University, China
{chengpang, zhangjie827}@sz.pku.edu.cn, hongliu@pku.edu.cn

Abstract

Direction of Arrival (DOA) estimation is an important technique for speech perception and speaker localization, which has received much attention in recent years. However, most conventional methods concentrate on noisy environments, which leads to serious degradation of their performance in the presence of reverberation. To resolve this phenomenon, a novel approach based on reverberation weighting and noise error estimator by two microphones is proposed for robust DOA estimation in this paper. Firstly, the reverberations in received microphone signals are suppressed by late and early reverberation gains estimated by a spectral subtraction rule and the coherence of direct-arrival signals respectively. Then, the reverberation-suppressed signals are utilized to extract the time difference of arrival (TDOA) by the noise error estimator to reduce the affect of noise. At last, the final DOA is determined by the obtained TDOA combing with the geometry of the microphone array. The proposed method is evaluated in a simulated rectangular room with different levels of noise and reverberation, and experimental results validate that our method achieves favorable performance compared with traditional ones.

Index Terms: Direction of Arrival, TDOA, reverberation weighting, noise error estimator

1. Introduction

Direction of Arrival estimation of speech signals has gained great interests in many applications such as hearing-aid, telephone communication and speaker localization in video conference [1–3]. Nevertheless, it still remains challenging to achieve robust DOA estimation due to the affect of noise, reverberation and other interferences.

As to DOA estimation, existing methods mainly concentrate on noisy environments [4–6], which can be divided into two groups: single- and dual-step approaches. Time difference of arrival (TDOA) based estimation is the classical method in the dual-step approaches. TDOA is firstly estimated from the microphone arrays, then the DOA of the sound is decided by the TDOA. For the single-step approaches, they can be further classified into two categories: the high-resolution spectral estimation methods (e.g. multiple signal classification, MUSIC) [7] and maximum-likelihood (ML) methods [8], such as steered response power (SRP) of beamformer [9]. The MUSIC method is based on subband processing, which achieves good performance for the narrow band signals. However, its performance degrades seriously for the broadband signals. Differently, SRP methods estimate the direction through maximizing of the output power of a beamformer to potential direction. SRP-PHAT [10], which is a typical version of SRP, has obtained a good performance in the noisy environments, but it needs more microphones such that it costs high computational complexity. For TDOA-based methods, TDOA is estimated by the spatially separated microphone pairs in the first stage. The generalized cross correlation (GCC) proposed by Knapp et al. [11] is commonly used for TDOA estimation. However, the GCC method is based on a reverberant-free model, it could not work in the reverberant environments. Although a cepstral prefiltering is proposed for GCC method (GCC-CEP) to reduce the influence of reverberation on TDOA estimation [12], whereas it looses effectiveness in the noisy environments.

Accordingly, this paper introduces a weighting function to eliminate reverberation and a noise error estimator to calculate TDOA. As to reverberation, it can be decomposed into early and late reverberation [13], which affects the sound differently. Since the late reverberation smears the signal, it brings the distortion of the available auditory cues for the DOA estimation. Hence, the late reverberation is reduced by a spectral subtraction rule as it can be considered as an uncorrelated noise process. The early reverberation brings about the confused peaks of cross correlation, because its energy is similar to the direct sound, so it is suppressed by attenuating all non-coherent parts. Then, the TDOA is estimated from the reverberation-suppressed signals through the noise error estimator. Finally, the TDOA is utilised to evaluate DOA based on the geometrical relation between microphones. Experimental results indicate that the proposed method achieves better performance than the conventional methods in both the noisy and reverberant environments.

The rest of the paper is organized as follows. The reverberation weighting and noise error estimator are introduced in Section 2. Section 3 gives the experimental setup and analysis. Finally, the conclusion of this paper is drawn in Section 4.

2. Direction of Arrival Estimation

This section mainly introduces the reverberation weighting and the extraction of TDOA for DOA estimation. Let $s(n)$ represent the sound source signal, the signals received by the two microphones $x_1(n)$ and $x_2(n)$ under the noisy and reverberant conditions can be model as

$$x_i(n) = h_i(n) * s(n) + v_i(n), \forall i = 1, 2, \tag{1}$$

where $h_i(n)$ and $v_i(n)$ denote the room impulse response (RIR) and additional noise, respectively. The framework of the proposed method is shown in Fig. 1. It is mainly constituted by two components including reverberation weighting and TDOA estimation based on noise error estimator.

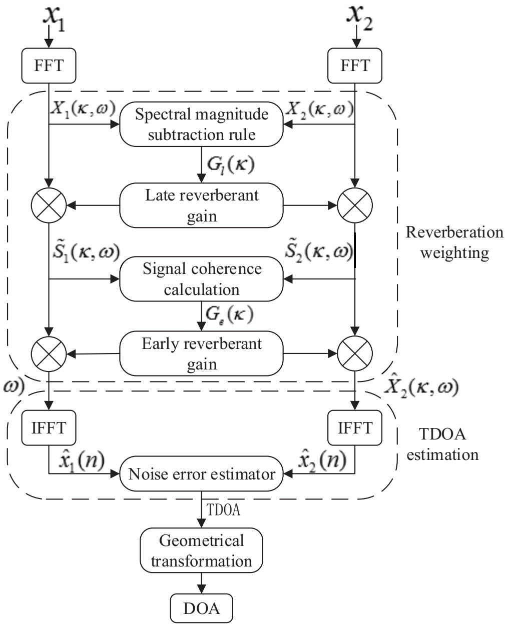

Figure 1: The framework of the proposed method consists of reverberation weighting and TDOA estimation.

2.1. Reverberation weighting

The room impulse response $h(n)$ consists of direct and early as well as late components, so it can be defined as

$$h(n) = \begin{cases} h_e(n), & 0 \le n < T_l \cdot f_s \\ h_l(n), & T_l \cdot f_s \le n \le T_r \cdot f_s \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where $h_e(n)$ concludes the direct and early propagating path of the sound source, $h_l(n)$ represents the late path, T_r refers to the reverberation time and f_s is the sampling frequency. T_l denotes the time span when late reverberation begins, which ranges from $50ms$ to $100ms$ [13]. Hence, the received signals can be denoted as

$$x_i(n) = \sum_{k=0}^{T_l f_s - 1} h_{i,e}(k)s(n-k) + \sum_{k=T_l f_s}^{T_r f_s} h_{i,l}(k)s(n-k) + v_i(n), \quad (3)$$

where $i = 1, 2$. Since the two components of room impulse response (RIR) affect the sound signal in different ways, they are treated separately as follows.

Here, reverberation is reduced by suppressing the late and early reverberation in the frequency domain using the late and early reverberation gains, which are calculated based on a spectral subtraction rule and the coherence of signals, respectively. In order to keep the TDOA between microphone signals unaffected, the same weighting gains are applied to each microphone signal. Since the same spectral weighting has no influence on the coherence of signals, the first step is to suppress the late reverberant components. For this, the variances of late reverberation can be acquired by a simple statistical model for the RIR [14]

$$\tilde{h}_l(n) = m(n)e^{-\rho n f_s^{-1}}, n \ge 0, \quad (4)$$

where $m(n)$ is a sequence of random variables with zero mean following normal distribution. And ρ is the decay rate, which is related to the reverberation time T_R through

$$\rho = \frac{3\ln(10)}{T_R}, \quad (5)$$

where T_R is estimated by the method proposed by Schroeder [15]. From Eq. (3), the late reverberant component can be considered as an uncorrelated noise process if the energy of direct path is smaller than all reflections [16].

So the variance of the late reverberant speech signal can be estimated by an estimator proposed by [17]

$$\sigma^2_{x_l}(\kappa,\omega) = e^{-2\rho T_l} \cdot \sigma^2_{x_l}(\kappa - N_l, \omega), \quad (6)$$

where $\sigma^2_{x_l}(\kappa,\omega)$ denotes the variance of the reverberant signal, N_l is the number of frames corresponding to T_l and κ is the frame index. Then, the spectral variance of the reverberant speech signal is calculated using recursive averaging

$$\sigma^2_{x_l}(\kappa,\omega) = \alpha_1 \cdot \sigma^2_{x_l}(\kappa - 1,\omega) + (1-\alpha_1)|X_l(\kappa,\omega)|^2, \quad (7)$$

where α_1 is a smoothing factor which ranges from 0 to 1, X_l is the late reverberant signal in the frequency domain. Then, a posteriori signal-to-interference ratio (SIR) can be obtained by

$$\eta(\kappa,\omega) = \frac{|X_l(\kappa,\omega)|^2}{\sigma^2_{x_l}(\kappa,\omega)}. \quad (8)$$

The weighting gain for suppressing the late reverberant components of microphone signal is calculated based on the spectral magnitude subtraction rule as

$$G_l(\kappa,\omega) = 1 - \frac{1}{\sqrt{\eta(\kappa,\omega)}}. \quad (9)$$

Hence, the signals whose late reverberations is suppressed can be derived from

$$\tilde{S}_i(\kappa,\omega) = \mathrm{X}_i(\kappa,\omega) \cdot G_l(\kappa,\omega), \forall i = 1, 2. \quad (10)$$

The second step is to achieve the suppression of the early reverberant components. Motivated by this, the coherence-based method is used to keep the coherent parts unaffected and remove the all non-coherent signal parts, because the direct-arrival signal shows a high coherence among different microphones. Here, the directly estimated coherence is used as weighting gain to suppress the early reverberation, it is defined as

$$G_e(\kappa,\omega) = \frac{|\Phi_{x_1x_2}(\kappa,\omega)|}{\sqrt{\Phi_{x_1x_1}(\kappa,\omega)\Phi_{x_2x_2}(\kappa,\omega)}}, \quad (11)$$

where $\Phi_{x_1x_2}(\kappa,\omega)$ and $\Phi_{x_ix_i}(\kappa,\omega), \forall i = 1, 2$ refer to the weighted short-term cross-correlation and auto-correlation functions, respectively, which can be evaluated as

$$\Phi_{x_ix_i}(\kappa,\omega) = \alpha_2\Phi_{x_ix_i}(\kappa-1,\omega) + |\tilde{S}_i(\kappa,\omega)|^2, \forall i = 1, 2, \quad (12)$$

$$\Phi_{x_1x_2}(\kappa,\omega) = \alpha_2\Phi_{x_1x_2}(\kappa-1,\omega) + \tilde{S}_1(\kappa,\omega)\tilde{S}^*_2(\kappa,\omega), \quad (13)$$

where $*$ denotes the complex conjugate and α_2 is a recursion factor, which determines the temporal integration time t of the coherence estimate. The relationship between α_2 and t is given by [18]

$$\alpha_2 = e^{-\frac{L}{4tf_s}}, \quad (14)$$

where L is the frame length.

In consequence, applying the early reverberant gains to $\tilde{S}_i(\kappa,\omega)$, we can get the output speech as

$$\hat{X}_i(\kappa,\omega) = \tilde{S}_i(\kappa,\omega) \cdot G_e(\kappa,\omega), \forall i = 1, 2. \quad (15)$$

The spectrograms of the reverberation-suppressed results are shown in Fig. 2 with $T_r = 0.5s$. It can be seen that the speech spectrum of the reverberated signal is smeared compared with the original signal and the tail of the reverberation is effectively reduced by the reverberation weighting. At last, the outputs $\hat{x}_i(n), i = 1, 2$ in the time domain can be obtained by the inverse discrete Fourier transform.

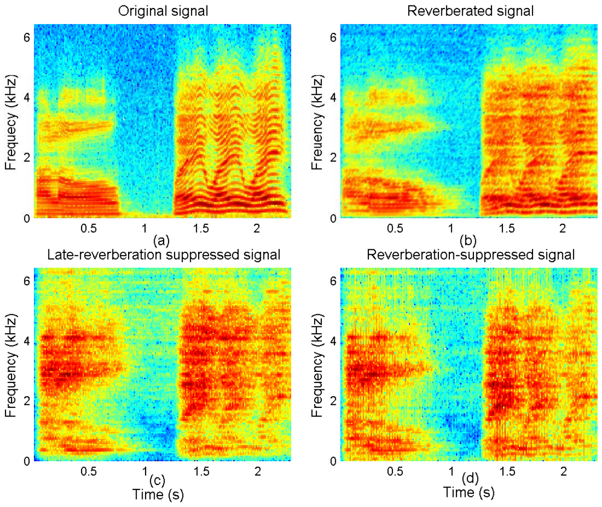

Figure 2: The spectrograms from original signal to reverberation weighted signal with $T_r = 0.5s$.

2.2. TDOA estimation based on noise error estimator

From the above reverberant weighting method, the resulting signal $\hat{x}_1(n)$, $\hat{x}_2(n)$ can be considered to only contain the direct-arrival signal and additional noises. Hence, $\hat{x}_1(n)$ and $\hat{x}_2(n)$ can be modeled as

$$\hat{x}_i(n) = a_i s(n - \tau_i) + v_i(n), \forall i = 1, 2, \tag{16}$$

where a_i represents the attenuation factors. Then, the Eq. (16) in the frequency domain can be shown as

$$\hat{X}_i(\omega) = a_i S(\omega) e^{-j\omega\tau_i} + V_i(\omega), \forall i = 1, 2. \tag{17}$$

Accordingly, the Eq. (17) can be transformed into

$$\frac{a_1}{a_2} e^{-j\omega\Delta\tau} = \frac{\hat{X}_1(\omega) - V_1(\omega)}{\hat{X}_2(\omega) - V_2(\omega)}, \tag{18}$$

where $\Delta\tau = \tau_1 - \tau_2$, which corresponds to the TDOA. Let $\gamma = a_1/a_2$, then the noise error estimator is defined as

$$\Delta V(\omega) = V_1(\omega) - \gamma V_2(\omega) e^{-j\omega\Delta\tau} = \hat{X}_1(\omega) - \gamma \hat{X}_2(\omega) e^{-j\omega\Delta\tau}. \tag{19}$$

Under the indoor environments, $\Delta V(\omega)$ is assumed to follow the zero-mean Gaussian distribution. Hence, the variance of $\Delta V(\omega)$ can be obtained by

$$Z(\omega) = \left\| \hat{X}_1(\omega) - \gamma \hat{X}_2(\omega) e^{-j\omega\Delta\tau} \right\|^2, \tag{20}$$

where $Z(\omega)$ and $\hat{X}(\omega)$ denote the Fourier transform of variance and the reverberation-suppressed signals, respectively. And let

$$Y(\omega) = \hat{X}_1(\omega) - \gamma \hat{X}_2(\omega) e^{-j\omega\Delta\tau}, \tag{21}$$

then, the result of $\partial Z(\omega)/\partial\Delta\tau$ can be calculated by

$$\frac{\partial Z(\omega)}{\partial \Delta\tau} = -2j\gamma\omega \hat{X}_2^*(\omega) Y(\omega) e^{-j\omega\Delta\tau}. \tag{22}$$

Let $\partial Z(\omega)/\partial\Delta\tau = 0$, since $j\omega$, γ and $e^{-j\omega\Delta\tau}$ are not equal to 0, it can be got

$$\hat{X}_2^*(\omega) \left(\hat{X}_1(\omega) - \gamma \hat{X}_2(\omega) e^{-j\omega\Delta\tau} \right) = 0, \tag{23}$$

then, we transform Eq.(23) into the time domain through the inverse Fourier transform, it can be given as

$$\delta(\tau - \Delta\tau) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\hat{X}_1(\omega)\hat{X}_2^*(\omega)}{\gamma \hat{X}_2(\omega)\hat{X}_2^*(\omega)} e^{j\omega\tau} d\omega. \tag{24}$$

Therefore, the time delay $\Delta\tau$ can be obtained from

$$\Delta\tau = \arg\max_{\tau} \frac{1}{2\pi\gamma} \int_{-\pi}^{\pi} \frac{\hat{X}_1(\omega)\hat{X}_2^*(\omega)}{\hat{X}_2(\omega)\hat{X}_2^*(\omega)} e^{j\omega n} d\omega, \tag{25}$$

since γ is a constant, so the value of $\Delta\tau$ is not affected by γ. Finally, the optimal time-delay $\Delta\tau$ is gained based on the Minimum Mean Square Error criterion.

2.3. DOA estimation

In this work, once TDOA is obtained, the corresponding DOA of the sound source is estimated based on the geometrical relationship between microphones:

$$\theta = \sin^{-1}\left(\frac{\Delta\tau c}{d f_s}\right), \tag{26}$$

where d refers to the distance between the two microphones, c represents the velocity of the sound in the air, which is usually set to $344m/s$.

3. Experiments and Discussions

3.1. Experimental environment and setup

The proposed method is evaluated in a rectangular room $(8 \times 4 \times 3)m$ simulated by the Roomsim toolbox [19], which is based on the image method [20]. The two microphones are placed at $(4, 1.93, 1.5)m$ and $(4, 2.07, 1.5)m$, respectively. The subject #21 in the CIPIC HRIR database is used as the Kemar head impulse response [21]. The detailed settings and parameters are shown in Fig. 3 and Table 1.

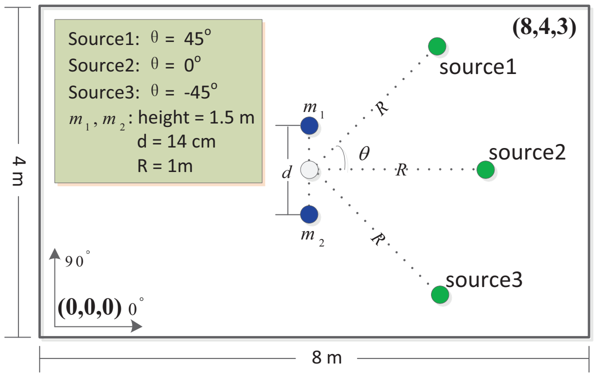

Figure 3: Simulation scene and parameters of the experimental environments.

The sound source is a musical period. The noise added to each microphone is zero-mean and independent Gaussian random noise, which is scaled to control the SNR. The proposed approach is evaluated in conditions with different Signal Noise Ratios (SNR) and reverberation times (T_r). The SNRs conclude 10dB and 40dB, which generally represent the office and quiet environment, respectively. And T_r varies form 0.1s to 0.5s. The proposed method is compared with the classical GCC-PHAT method [11], a modified GCC-ML method [8] and GCC method based on cepstral prefiltering (GCC-CEP) [12].

Table 1: Parameters used in experiments

Parameter	Value
Sampling frequency f_s	44.1kHz
Frame length (FFT length)	256 points
Frame overlap	128 points
Block length (observation time)	2s
smoothing factor α_1	0.95
recursion factor α_2	0.97

3.2. Experimental results and analysis

In this paper, the accuracy and Root Mean Square Error (RMSE) of the DOA estimation are utilised to verify the effectiveness of the proposed method. The DOA estimation is thought to be correct when the estimated DOA is located around the true DOA with a certain tolerance. So the accuracy of DOA P_a is given by:

$$P_a = N_c/N_t, \tag{27}$$

where N_c represents the correct DOA estimates and N_t is the total DOA estimates. Then, the RMSE is defined as [4]:

$$RMSE = \sqrt{\frac{1}{N_t}\sum_{n=1}^{N_t}(\theta - \hat{\theta})^2}, \tag{28}$$

where θ and $\hat{\theta}$ represent the true DOA and the estimated DOA, respectively.

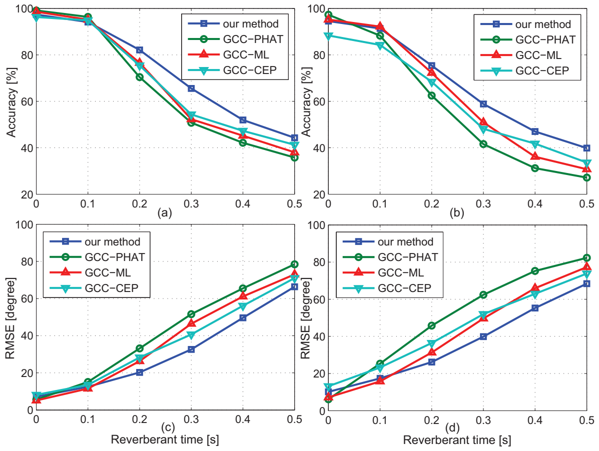

Figure 4: The accuracy and RMSE of DOA estimation at different reverberation times with SNR=40dB (left), SNR=10dB (right). The tolerance of accuracy is 5°.

When we place a speech source at $\theta = 0^\circ$ with different SNRs and reverberant times, the experimental results of our method compared with state-of-the-art methods are shown in Fig.4. As to the results, the accuracies among the four methods are similar with each other under the environment without reverberation and all the accuracies are over 90% with the tolerance 5°. Nevertheless, the proposed method outperforms the other methods with the increase of reverberation time, especially in the strong reverberant conditions. This superiority primarily owes to the usage of reverberation weighting, which extracts the direct signal to estimate TDOA through suppressing the late and early reverberation by a spectral subtraction rule and the coherence of signals, respectively. However, the performance of the proposed method is a little worse than the other three methods in the weak reverberant environment ($T_r = 0.1s$), which is due to that the reverberation weighting leads to a distortion of the original received signal. The performance of GCC-PHAT method declines severely with the extension of the reverberation time, because the cross correlation of the received signals is seriously disturbed by reverberation so that it generates a wrong peak. The GCC-ML method achieves a favorable performance in the weak reverberant environment because of its robustness to noise, while its performance drops sharply when $T_r \geq 0.3s$, because it ignores the early reverberation and takes the whole reverberation as noise, which results in serious distortion of the TDOA extraction. The GCC-CEP method behaves better than the other two GCC methods, but its performance decreases seriously when SNR becomes smaller, which is caused by neglecting the influence of noise. As for RMSE, it is obvious that our method achieves smaller RMSE than the other three methods overall, which owns to the noise error estimator reduces the noisy effect by minimizing the variance of noise. Along with the increase of reverberation time, the RMSEs of GCC-PHAT and GCC-ML are lager than others, which is attributed to ignoring the difference between the late and early reverberation. In the strong noisy environment (SNR=10dB), GCC-CEP obtains worse performance as it does not take the noise into account.

Table 2: The accuracy of DOA in different conditions.

SNR	40dB		10dB	
Tolerance	0^o	5^o	0^o	5^o
$T_r = 0.1s$	89.43%	96.13%	71.26%	83.37%
$T_r = 0.3s$	63.96%	69.53%	52.72%	57.39%
$T_r = 0.5s$	37.26%	44.29%	29.74%	38.13%

The average accuracies of the proposed method under different conditions are illustrated in Table 2. It can be observed from Table 2 that the accuracy of the proposed method has exceeded 50% in mild reverberant environments ($T_r \leq 0.3s$). Above all, our method is more available and practical for DOA estimation compared with other methods.

4. Conclusions

In this work, a new and effective DOA estimation method using the reverberation weighting and noise error estimator is proposed for the noisy and reverberant conditions. The reverberation weighting relies on the different influence of late and early reverberation gains in the frequency domain, which can keep the original time difference information unaffected while suppressing the reverberation. Noise error estimator is involved to time delay estimation by minimizing the variance of noise, which can decrease the error of TDOA estimation. The effectiveness of the proposed method is evaluated in the simulated experiments, which proves that our method is more suitable for real environments. However, the performance the proposed method suffers little degradation in the weak reverberant environments. Our future work will concentrate on the influence of reverberation weighting on the direct signals.

5. Acknowledgements

This work is supported by National Natural Science Foundation of China (NSFC, No.60875050, 60675025, 61340046), National High Technology Research and Development Program of China (863 Program, No.2006AA04Z247), Science and Technology Innovation Commission of Shenzhen Municipality (No.201005280682A, No.JCYJ20120614152234873), Specialized Research Fund for the Doctoral Program of Higher Education (No. 20130001110011).

6. References

[1] W. Xue, S. Liang, and W. Liu, "Weighted spatial bispectrum correlation matrix for doa estimation in the presence of interferences," in *INTERSPEECH*, pp. 2228–2232, 2014.

[2] H. Liu, J. Zhang, and Z. Fu, "A new hierarchical binaural sound source localization method based on interaural matching filter," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1598–1605, 2014.

[3] X. Li and H. Liu, "Sound source localization for hri using foc-based time difference feature and spatial grid matching," *IEEE Transactions on Cybernetics*, vol. 43, no. 4, pp. 1199–1212, 2013.

[4] W. Xue and W. Liu, "Direction of arrival estimation based on subband weighting for noisy conditions," in *INTERSPEECH*, pp. 142–145, 2012.

[5] H. Liu and J. Zhang, "A binaural sound source localization model based on time-delay compensation and interaural coherence," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1424–1428, 2014.

[6] H. Liu and X. Li, "Time delay estimation for speech signal based on foc-spectrum," in *INTERSPEECH*, pp. 1732–1735, 2012.

[7] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[8] Y. Rui and D. Florencio, "Time delay estimation in the presence of correlated noise and reverberation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 133–136, 2004.

[9] J. M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach," *IEEE International Conference on Robotics and Automation*, vol. 1, pp. 1033–1038, 2004.

[10] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer, 2001.

[11] C. 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.

[12] A. Stéphenne and B. Champagne, "Cepstral prefiltering for time delay estimation in reverberant environments," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 3055–3058, 1995.

[13] J. Blauert, *The technology of binaural listening*. Springer, 2013.

[14] K. Lebart, J. M. Boucher, and P. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[15] M. R. Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, no. 3, pp. 409–412, 1965.

[16] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," *Dissertation Abstracts International*, vol. 68, no. 04, 2007.

[17] M. Jeub, M. Schafer, T. Esch, and P. Vary, "Model-based dereverberation preserving binaural cues," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1732–1745, 2010.

[18] A. Westermann, J. M. Buchholz, and T. Dau, "Binaural dereverberation based on interaural coherence histogramsa," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 2767–2777, 2013.

[19] D. Campbell, K. Palomaki, and G. Brown, "A matlab simulation of shoebox room acoustics for use in research and teaching," *Computing and Information Systems*, vol. 9, no. 3, pp. 48–51, 2005.

[20] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[21] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The cipic hrtf database," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp. 99–102, 2001.