DOA ESTIMATION OF SPEECH SOURCE IN NOISY ENVIRONMENTS WITH WEIGHTED SPATIAL BISPECTRUM CORRELATION MATRIX

Wei Xue, Shan Liang, Wenju Liu

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

ABSTRACT

Although the high order statistics (HOS) has promising property against the Gaussian noise, there still lack effective ways to apply the HOS to DOA estimation of the speech source. In this paper, we propose a novel HOS based DOA estimation method for speech source in strong noise conditions. A "weighted spatial bispectrum correlation matrix (WSBCM)" is formulated, which contains the spatial correlation information of bispectrum phase differences. We then propose a new DOA estimator based on the eigenvalue analysis of the WSBCM. Besides the theoretical advantage of the bispectrum against Gaussian noises, the redundant information in the bispectrum domain is also exploited to make the WSBCM noise robust. The WSBCM enables bispectrum weighting to select the speech units in the bispectrum, which further helps to improve the performance. Experimental results demonstrate that the proposed method outperforms existing algorithms in different kinds of noisy environments.

Index Terms— direction of arrival estimation, microphone array signal processing, bispectrum

1. INTRODUCTION

Direction of arrival (DOA) estimation, which aims at determining the direction of sound sources using the microphone array, has drawn much attention due to its wide applications such as in interactive robots, video conferences, and hand free devices. In practical scenarios, one critical issue is to make the DOA estimator robust when the strong noise exists in the environment.

Conventional DOA estimation methods can be generally classified into three categories: high-resolution spectral estimation [1, 2, 3], steered beamformer response power [4, 5], and time difference of arrival (TDOA) estimation [6, 7]. The classical methods can perform well in moderate spatially white Gaussian noises, however, when the noise level is high, they always suffer from severe performance degradation.

As the high order cumulant or spectrum of the Gaussian signal is always zero, some HOS based methods have been proposed to improve the performance against Gaussian noise [8, 9, 10, 11, 12]. However, almost all these algorithms are proposed for narrowband signals, in order to apply them for the broadband speech signal, it is much time-consuming if we decompose the speech signal into narrowband signals and estimate the DOAs separately in each narrowband. Moreover, the estimation accuracy of HOS drops significantly when the data length become short [13, 14], this negative effect further limits the application of HOS based methods on speech as the speech signal is only short-time stationary.

In order to better apply the HOS to the DOA estimation of speech source, we propose a novel method in this paper. A "WSBCM" is formulated, which contains the spatial correlation information of the phase differences in the bispectrum domain. The WSBCM is a function of a hypothesized DOA, and shows interesting property only when the hypothesized DOA equals to the true one. It avoids the separate narrowband estimation procedures, and exploits the redundant information in the bispetrum domain to reduce the effect of noise and bispectrum estimation error. In addition, the WSBCM enables bispectrum weighting to further improve the robustness to noise. We finally propose a new DOA estimator based on the eigenvalue analysis of the WSBCM. Experimental results demonstrate that the proposed method outperforms existing algorithms in different noisy environments.

2. PROBLEM FORMULATION

We consider the problem in an environment with an M-element uniform linear microphone array (ULA), a speech source and several Gaussian noise sources. All sound sources are in the far field [15] and uncorrelated with each other.

Assuming the signals are equally attenuated from the speech source to each microphone, the signal received by the mth microphone at time k can be simply expressed as:

$$y_m(k) = s(k - \tau_m) + v_m^C(k), m = 1, 2, ..., M, \quad (1)$$

where $s(k)$ is the speech signal, τ_m is the propagation time from the speech source to the mth microphone, and the $v_m^C(k)$ denotes the combined noise which consists of the directional and spatially white Gaussian noises.

This research was supported in part by the China National Nature Science Foundation (No.91120303, No.61273267 and No.90820011).

If we indicate the received speech signal $s(k-\tau_m)$ as $x_m(k)$, and choose the first microphone as reference, it can be easily seen that:

$$x_m(k) = x_1(k-\tau_{m1}), m = 1, 2, ..., M, \quad (2)$$

where τ_{m1} stands for the TDOA between the mth and first microphone. For ULA, according to the array geometry, τ_{m1} can be derived as follows:

$$\tau_{m1} = \tau_m - \tau_1 = (m-1)\frac{\sin(\hat{\theta}) f_s d}{c}, m = 1, 2, ..., M, \quad (3)$$

where c is the speed of sound in the air, f_s is the sampling rate, d is the spacing between two adjacent microphones and $\hat{\theta}$ is the true DOA to be estimated.

3. PROPOSED METHOD

3.1. Phase Difference in the Bispectrum Domain

As most speech signals have asymmetric pdf's, their skewness are non-zero [14]. Therefore, it is reasonable to analyze the speech signal in the bispectrum domain.

Recall the signals received by first and mth microphone. Following Eq.(1)$\sim$(3), they can be rewritten as follows:

$$\begin{aligned} y_1(k) &= x_1(k) + v_1^C(k) \\ y_m(k) &= x_1(k - \frac{(m-1)\sin(\hat{\theta}) f_s d}{c}) + v_m^C(k). \end{aligned} \quad (4)$$

Since $v_i^G(k)$, for $i = 1, m$, is zero-mean Gaussian, its bispectrum is identical to zero. According to the derivation in [16], the following relationships hold for the bispectrum of $y_1(k)$ and the cross-bispectrum between $y_1(k)$, $y_m(k)$:

$$\begin{aligned} B_{y_1 y_1 y_1}(\Omega_1, \Omega_2) &= B_{x_1 x_1 x_1}(\Omega_1, \Omega_2) \\ B_{y_1 y_m y_1}(\Omega_1, \Omega_2) &= B_{x_1 x_1 x_1}(\Omega_1, \Omega_2) e^{j\Omega_1 \frac{(m-1)\sin(\hat{\theta}) f_s d}{c}}, \end{aligned} \quad (5)$$

where $B_{xyz}(\Omega_1, \Omega_2)$ stands for the bispectrum of signal $x(k)$,$y(k)$ and $z(k)$, Ω_1 and Ω_2 are the bi-frequencies. We define the bispectrum phase difference (BPD) as the ratio between $B_{y_1 y_m y_1}$ and $B_{y_1 y_1 y_1}$:

$$I_{m1}(\Omega_1, \Omega_2) \stackrel{\text{def}}{=} \frac{B_{y_1 y_m y_1}(\Omega_1, \Omega_2)}{B_{y_1 y_1 y_1}(\Omega_1, \Omega_2)} = e^{j\Omega_1 \frac{(m-1)\sin(\hat{\theta}) f_s d}{c}}. \quad (6)$$

It can be seen from Eq.(6) that, in theory, the effect of Gaussian noise has been fully removed in BPD. When multiple microphones are available, one may solve the narrowband DOA estimation problem using methods such as MUSIC in each (Ω_1, Ω_2) unit, by taking the values of $I_{m1}(\Omega_1, \Omega_2)$ for $m = 1, ..., M$ as the multichannel narrowband signals, and finally combine the estimation results of all (Ω_1, Ω_2) units. However, it is computationally expensive. Moreover, although the BPD seems promising theoretically against the Gaussian noise, in practice, estimating the bispectrums from short signal sequences causes large error, which limits the performance of bispectrum based DOA estimation.

3.2. Weighted Spatial Bispectrum Correlation Matrix

In order to cope with the problems described above, in this subsection, we formulate a new matrix called "WSBCM" which will be used by the DOA estimator in the next subsection. The WSBCM reflects the spatial correlations between the phase aligned multichannel BPDs for a hypothesized DOA, and shows interesting property only when the hypothesized DOA equals to the true one. It avoids estimating the DOA separately in each (Ω_1, Ω_2) unit, and exploits the redundant information of BPDs in the bi-frequency Ω_2 to reduce the negative effect of noise and large bispectrum estimation error. In addition, the WSBCM allows the bispectrum weighting to further improve the robustness to noise.

For a certain value of Ω_1 and Ω_2, we define the BPD vector as:

$$\mathbf{I}(\Omega_1, \Omega_2) \stackrel{\text{def}}{=} [I_{11}(\Omega_1, \Omega_2), \ldots, I_{M1}(\Omega_1, \Omega_2)]^T. \quad (7)$$

The vector consists of the BPDs of all microphones in the (Ω_1, Ω_2) unit. Obviously, an explicit theoretical expression of the BPD vector can be derived according to Eq.(6), therefore, if we know the true DOA exactly, the BPDs can be totally compensated by a corresponding "BPD compensation vector". Here a BPD compensation vector for a hypothesized DOA θ is defined as:

$$\mathbf{C}(\theta, \Omega_1) \stackrel{\text{def}}{=} [1, e^{-j\Omega_1 \frac{\sin(\theta) f_s d}{c}}, \ldots, e^{-j\Omega_1 \frac{(M-1)\sin(\theta) f_s d}{c}}]^T. \quad (8)$$

Then we compensate the BPDs using the compensation vector as follows:

$$\mathbf{I}^C(\theta, \Omega_1, \Omega_2) \stackrel{\text{def}}{=} \mathbf{I}(\Omega_1, \Omega_2) \circ \mathbf{C}(\theta, \Omega_1), \quad (9)$$

where $\mathbf{I}^C(\theta, \Omega_1, \Omega_2)$ is the phase-compensated BPD vector, and the symbol "$\circ$" stands for the Hadamard product. Once θ is equal to $\hat{\theta}$, according to Eq.(6)(8)(9),

$$\mathbf{I}^C(\theta, \Omega_1, \Omega_2) = \mathbf{\Gamma} + \mathbf{E}(\theta, \Omega_1, \Omega_2), \quad (10)$$

where $\mathbf{\Gamma} = [1, 1, \ldots, 1]^T$, which indicates that BPDs are phase aligned, and $\mathbf{E}(\theta, \Omega_1, \Omega_2)$ is an error term. As is illustrated in Fig.1, the bispectrum of the speech signal does not distribute uniformly in all (Ω_1, Ω_2) units. In some speech-absent (Ω_1, Ω_2) units, the large BPD error (i.e. the error of the elements of $\mathbf{I}(\Omega_1, \Omega_2)$) exists, which will indirectly make the $\mathbf{E}(\theta, \Omega_1, \Omega_2)$ also large compared with $\mathbf{\Gamma}$. Therefore, it is better to consider only the "speech" units in the bispectrum.

We select the "speech" units by the bispectrum weight $w(\Omega_1, \Omega_2)$ with non-zero values only in the "speech" units. The $w(\Omega_1, \Omega_2)$ is simply defined according to the amplitude of the bispectrum in each unit:

$$w(\Omega_1, \Omega_2) \stackrel{\text{def}}{=} max(|B_{y_1 y_1 y_1}(\Omega_1, \Omega_2)| - \zeta, 0), \quad (11)$$

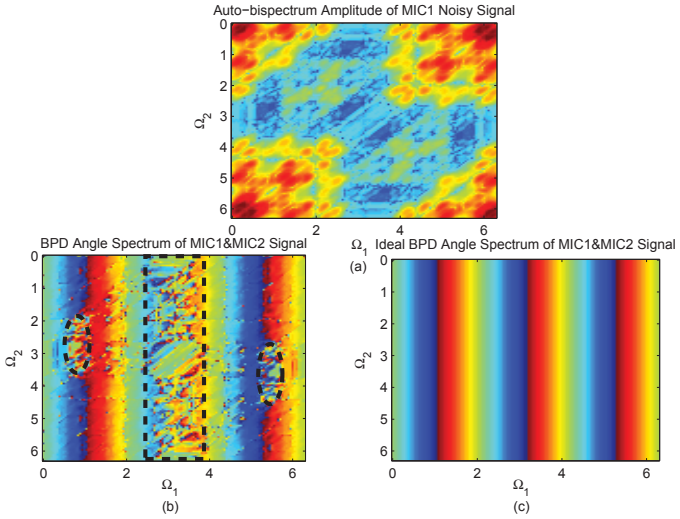

Fig. 1. *Example of the relationship between the BPD estimation error and the bispectrum amplitude of one speech frame. Compared with (c), the circled areas in (b) have large BPD estimation error. These areas correspond to the low-amplitude non-speech areas in (a).*

where $|\cdot|$ is the "taking the amplitude" operator, and ζ is a threshold. The weight of $w(\Omega_1, \Omega_2)$ is set to be zero if its bispectrum amplitude is lower than ζ.

With these bispectrum weights, we define the WSBCM which contains the spatial correlation information of the phase aligned BPDs for a hypothesized DOA as follows:

$$\mathbf{R}(\theta) \stackrel{\text{def}}{=} \sum_{\Omega_1,\Omega_2} w(\Omega_1, \Omega_2)[\mathbf{I}^C(\theta, \Omega_1, \Omega_2)][\mathbf{I}^C(\theta, \Omega_1, \Omega_2)]^H. \tag{12}$$

We assume that in the speech bispectrum units, the error term in Eq.(10) can be ignored. Then, once the hypothesized DOA θ is equal to $\hat{\theta}$, according to Eq.(10)(12),

$$\begin{aligned}\mathbf{R}(\theta) &= \sum_{(\Omega_1,\Omega_2)\in\Omega_S} w(\Omega_1, \Omega_2)\mathbf{\Gamma}\mathbf{\Gamma}^H \\ &= \eta \cdot \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix},\end{aligned} \tag{13}$$

where Ω_S denotes the set of speech units in the bispectrum which are selected by the bispectrum weight, and $\eta = \sum_{(\Omega_1,\Omega_2)\in\Omega_S} w(\Omega_1, \Omega_2)$ which is a constant in a certain frame given $w(\Omega_1, \Omega_2)$. In this case, $\mathbf{R}(\theta)$ is a matrix of rank 1. If $\theta \neq \hat{\theta}$, the $\mathbf{R}(\theta)$ will be semi-definite, and its rank will be greater than 1.

Now let's further analyze the definition of $\mathbf{R}(\theta)$ in Eq.(12). According to Eq.(6)∼(9), the factor $\mathbf{I}^C(\theta, \Omega_1, \Omega_2)$ in Eq.(12) is actually not a function of Ω_2 theoretically. Therefore, if ignoring the effect of $w(\Omega_1, \Omega_2)$, as the values of $\mathbf{I}^C(\theta, \Omega_1, \Omega_2)$ are repeated for different Ω_2's, it seems that summing over all Ω_2's brings no more information than just summing over an arbitrary fixed Ω_2. However, because the factor $\mathbf{I}(\Omega_1, \Omega_2)$ in $\mathbf{I}^C(\theta, \Omega_1, \Omega_2)$ is computed from the estimated bispectrums, in practice, these values of $\mathbf{I}^C(\theta, \Omega_1, \Omega_2)$ for different Ω_2's will not exactly stay unchanged. Then by summing over all Ω_2's, the redundant information in the bi-frequency Ω_2 brings more data to suppress the effect of noise and bispectrum estimation error, and to test the degree of alignment of phase compensated BPDs for a hypothesised DOA.

3.3. DOA estimator

In this subsection, we define a new DOA estimator based on the eigenvalue analysis of the WSBCM. Let us perform the eigenvalue decomposition of $\mathbf{R}(\theta)$ and let $\lambda_1(\theta) \geq \lambda_2(\theta) \geq \ldots \geq \lambda_N(\theta)$ denote the N eigenvalues of $\mathbf{R}(\theta)$. Obviously, if the hypothesized DOA θ equals to $\hat{\theta}$, $\mathbf{R}(\theta)$ is of rank 1, $\lambda_2(\theta) = \ldots = \lambda_N(\theta) = 0$. Therefore, if we form the following cost function

$$\mathbf{J}(\theta) \stackrel{\text{def}}{=} \frac{1}{\sum_{i=2}^{N} |\lambda_i(\theta)|}, \tag{14}$$

the cost function reaches the maximum if $\theta = \hat{\theta}$. Then the estimated DOA $\tilde{\theta}$ is calculated as:

$$\tilde{\theta} \stackrel{\text{def}}{=} arg \max_{\theta} \mathbf{J}(\theta). \tag{15}$$

4. EXPERIMENT

As is mentioned before, although some HOS based methods have been proposed, they are designed either for narrow-band signals or for long data sequences, and could not be used directly for speech source. So the proposed algorithm is compared with the well-known SRP-PHAT [5] and broad-band MUSIC algorithm [3] in spatially white and directional Gaussian noise conditions under different signal to noise ratio (SNR)s.

4.1. Experimental setup and evaluation

A rectangular room with size $6 \times 4 \times 3$ meters is modeled in the experiment. We employ a ULA which consists of eight omni-directional microphones, with the spacing between adjacent microphones as 10 cm. The microphones at two ends of ULA are at (2.5,2.0,1.5), (3.2,2.0,1.5) respectively. The speech source is located on a horizontal plane (x,y,1.5) with distance 2m to the center of the ULA. In order to facilitate the test, We only consider DOAs of the speech source ranging from $-90°$ to $90°$ with a step size of $20°$. For directional noise cases, three possible DOAs of the noise source are considered, which are $20°$, $40°$ and $60°$ respectively.

The room impulse response from the source to each microphone is modeled by image-source method [17]. In the experiment, we set the reverberant time T60 to be 250ms. The speech source is of 10 seconds, sampled with 16 bit resolution and 8KHz sampling rate. The received speech signal and noise signal (directional / spatially white) are separately generated, and mixed together after being scaled to control the SNR. The SNR changes from -10dB to 20dB, with a step size of 5dB. For all evaluated algorithms, the frame size is set to be 512 samples with 50% overlap. 50 Monte Carlo simulations are conducted for each scenario (noise type, SNR).

We use two frame level metrics, denoted as Accuracy and Root Mean Square Error (RMSE), to evaluate the performance of different algorithms. The estimation is considered to be correct if $|\tilde{\theta} - \hat{\theta}| < Th$, where Th is a threshold which is commonly set to be 5°. Then Accuracy and RMSE are defined as:

$$Accuracy \stackrel{\text{def}}{=} N_c/N, RMSE \stackrel{\text{def}}{=} \sqrt{E\{(\tilde{\theta} - \hat{\theta})^2\}}, \quad (16)$$

where N_c is the number of speech frames which have the correct estimation, N is the number of total speech frames. We only consider the speech frames for evaluation, and the speech frames are labeled manually in advance on the clean speech signal. It should be pointed out that these labels are never used by any of the three algorithms.

4.2. Experimental results

As are shown in Fig.2(a) and Fig.2(c), in the spatially white Gaussian noise conditions, the proposed algorithm yields the highest estimation accuracy in all SNRs considered, and gets the lowest RMSE at the same time. Even when SNR = -10dB, the proposed method can still achieve the estimation accuracy higher than 90%.

Similar comparison results can be observed in Fig.2(b) and Fig.2(d) for the directional Gaussian noise cases. Obviously, all the three algorithms suffer from performance degradation compared with the spatially white Gaussian noise cases in low SNR conditions, but the proposed algorithm degrades least. It should not come as a surprise that the SRP-PHAT algorithm totally breaks down under the strong directional noise. After performing the phase transform (PHAT) which discards the amplitudes of the cross-spectrums, all the frequency bins are treated with equal significance. As the speech signal occupies only a few frequency bins, when the strong directional noise source exists, most frequency bins are dominated by the noise source rather than the speech one. As a result, the DOA estimator which exploits the information of all frequency bins (integration or summation) will finally direct its global peak towards the noise source direction. The broadband MUSIC improves the performance to some extent, nevertheless, the proposed method achieves the best result. It can be seen that when the SNR is higher than 0 dB, the proposed method can perform quite reliably.

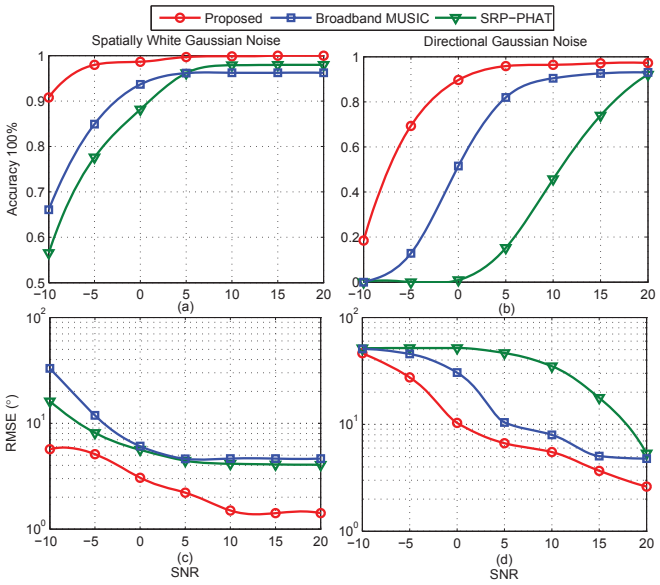

Fig. 2. *Estimation performance of different algorithms. (a)(c) Accuracy and RMSE under spatially white Gaussian noise. (b)(d) Accuracy and RMSE under directional Gaussian noise. The error tolerance for Accuracy is* 5°

5. CONCLUSION

In this paper, a new DOA estimation method for speech source in noisy conditions is proposed. The method is based on the "WSBCM", which exploits the redundant information of bispectrum phase differences to reduce the effect of bispectrum estimation error for short data sequences and avoids the separate narrowband DOA estimations. In addition, the WSBCM enables bispectrum weighting to further improve the robustness to noise. We also propose a new DOA estimator based on the eigenvalue analysis of the WSBCM. Experimental results show that the proposed method can achieve better performance than existing methods in different noisy environments.

6. RELATION TO PRIOR WORK

The theoretical basis of the BPD between two signals under the Gaussian noise in Eq.(5) can be found in [16]. In that paper, the authors proposed a method to estimate the time delay between two signals by bispectrum analysis. However, it gave only one estimation for the whole signal sequence, and how to cope with the bispectrum estimation error for short frames was not addressed. Moreover, it was not straightforward to extend the method to the multi-sensor cases. In this paper, based on the theoretical foundation of BPD, we concentrate more on how to exploit multiple microphones in a proper way to improve the DOA performance, and how to make the bispectrum based method applicable to short speech frames.

7. REFERENCES

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

[2] Michael L McCloud and Louis L Scharf, "A new subspace identification algorithm for high-resolution doa estimation," *Antennas and Propagation, IEEE Transactions on*, vol. 50, no. 10, pp. 1382–1390, 2002.

[3] Jacek P Dmochowski, Jacob Benesty, and Sofiene Affes, "Broadband music: opportunities and challenges for multiple source localization," in *Applications of Signal Processing to Audio and Acoustics, IEEE Workshop on*. IEEE, 2007, pp. 18–21.

[4] Jacek P Dmochowski, Jacob Benesty, and Sofiene Affes, "A generalized steered response power method for computationally viable source localization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2510–2526, 2007.

[5] M. Brandstein and D. Ward, "Microphone arrays: Signal processing techniques and applications," *Springer*, 2010.

[6] Tsvi G Dvorkind and Sharon Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Processing*, vol. 85, no. 1, pp. 177–204, 2005.

[7] Michael S Brandstein, John E Adcock, and Harvey F Silverman, "A practical time-delay estimator for localizing speech sources with a microphone array," *Computer Speech & Language*, vol. 9, no. 2, pp. 153–169, 1995.

[8] Philippe Forster and Chrysostomos L Nikias, "Bearing estimation in the bispectrum domain," *Signal Processing, IEEE Transactions on*, vol. 39, no. 9, pp. 1994–2006, 1991.

[9] Zhenghao Shi and Frederick W Fairman, "A comprehensive approach to doa estimation using higher-order statistics," *Circuits, Systems and Signal Processing*, vol. 17, no. 4, pp. 539–557, 1998.

[10] Zhenghao Shi and Frederick W Fairman, "Doa estimation via higher-order cumulants: a generalized approach," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 5, pp. 209–212, 1992.

[11] Boaz Porat and Benjamin Friedlander, "Direction finding algorithms based on high-order statistics," *Signal Processing, IEEE Transactions on*, vol. 39, no. 9, pp. 2016–2024, 1991.

[12] Norman Yuen and Benjamin Friedlander, "Doa estimation in multipath: an approach using fourth-order cumulants," *Signal Processing, IEEE Transactions on*, vol. 45, no. 5, pp. 1253–1263, 1997.

[13] Ines Jebali Gdoura, P Loizou, and Andreas Spanias, "Speech processing using higher order statistics," in *Circuits and Systems, IEEE International Symposium on*, 1993, pp. 160–163.

[14] JWA Fackrell and S McLaughlin, "The higher-order statistics of speech signals," in *Techniques for Speech Processing and their Application, IEE Colloquium on*. IET, 1994, pp. 7–11.

[15] Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang, "Springer handbook of speech processing," *Springer*, 2008.

[16] Jitendra K. Tugnait, "Time delay estimation with unknown spatially correlated gaussian noise," *Signal Processing, IEEE Transactions on*, vol. 41, no. 2, pp. 549–558, 1993.

[17] Eric A Lehmann and Anders 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, pp. 269, 2008.