TOWARDS ROBUST CLOSE-TALKING MICROPHONE ARRAYS FOR NOISE REDUCTION IN MOBILE PHONES

Edwin Mabande, Fabian Kuech, Alexander Niederleitner, and Anthony Lombard

Fraunhofer IIS, Am Wolfsmantel 33, D-91058 Erlangen, Germany
edwin.mabande@iis.fraunhofer.de

ABSTRACT

Adaptive close-talking differential microphone arrays (ACTMAs) inherently suppress farfield noise while emphasizing desired nearfield signals. This paper discusses the applicability of ACTMAs for noise reduction in mobile phones. In order to utilize the advantages of ACTMAs, we need to improve the robustness to microphone mismatch and improve parameter estimation accuracy. In this paper we propose a method to improve the robustness of the ACTMA algorithm by taking microphone gain mismatch into account in the detection of background noise and mobile phone user activity, performing online microphone gain calibration, steering the null of the ACTMA to the rear of the mobile phone, and performing parameter estimation only when mobile phone user activity is detected. Thus, the robust ACTMA is applicable for performing noise reduction in mobile phones. Experiments with recorded data demonstrate the effectiveness of this method.

Index Terms— Noise reduction, Adaptive close-talking microphone array, Mobile phone

1. INTRODUCTION

Mobile phones are used for telecommunication in widely differing acoustic environments. However, if conversations take place in adverse acoustical environments, i.e., high background noise, this may lead to a significant degradation of speech intelligibility and listening comfort for the listener at the far-end [1]. In such scenarios, the application of noise reduction algorithms [2, 3, 4] that ensure minimal speech distortion is highly desirable. Most of the mobile phones nowadays have two or more microphones and it has been shown that the noise reduction performance can be enhanced by exploiting the additional spatial diversity [1, 4].

In this paper, we discuss the application of adaptive close-talking differential microphone arrays (ACTMAs) [5, 6] for noise reduction in mobile phones. A prerequisite for the application of ACTMAs is the existence of two closely-spaced microphones. A common microphone configuration found in mobile phones is one in which there is a microphone at the bottom and another at the top of the mobile phone. Due to the small sizes of the MEMS (MicroElectrical-Mechanical System) microphones typically used in mobile phones, it becomes feasible to place an additional microphone at the bottom of the mobile phone in the configuration depicted in Figure 1. Note that the axis of the two-element array, consisting of microphones m_1 and m_2, is perpendicular to the front of the phone, i.e., the user is typically located at endfire. This configuration is chosen because higher gain is achieved at endfire [7, 8].

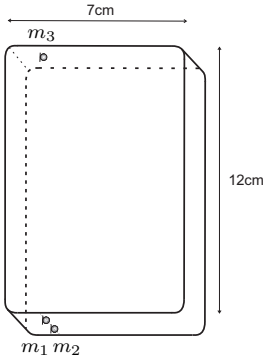

Fig. 1. Mobile phone illustration with three microphones, i.e., one at the top and two at the bottom.

There are two main challenges in the application of the ACTMAs in the mobile phone scenario; Microphone mismatches cause a significant degradation in the performance of the ACTMA algorithm. This necessitates a calibration of the microphones, which typically cannot be performed offline. In order to ensure the desired signal is not distorted, a correction filter [5] has to be computed based on the estimated positional information of the mobile phone user. To ensure sufficient accuracy, the estimation of the positional information should only occur during speech activity of the mobile phone user. Therefore, a method to detect the presence of speech from the mobile phone user is required.

In this paper, we show that by exploiting normalized power level differences (NPLDs) [4], we can overcome these challenges. We also show that for real measurements, microphone gain mismatches result in biased NPLDs measurements. We therefore propose the use of an adaptive threshold to improve robustness. In addition, it is necessary to steer the null of the ACTMA towards an angular region which does not overlap with the angular region in which the mobile phone user is typically found.

2. ACTMA

In the following, the ACTMA [5] is briefly described. Here, we assume a free-field model and that the mobile phone user's mouth is located close to the two microphones while the interfering sources are assumed to be far away. The ACTMA depicted in Figure 2 constitutes a first-order close-talking differential microphone array (CTMA), consisting of two closely-spaced omnidirectional ele-

This work was partially supported by the *FuE-Programm "Mikrosystemtechnick Bayern"des Bayerischen Staatsministeriums für Wirtschaft und Medien, Energie und Technologie (StMWMET)* within the twinMikro Project under project number BAY176/003.

ments, whose output is processed by an adaptive correction filter. Here, d is the distance between the microphones and θ_s is the desired source's direction of arrival (DOA).

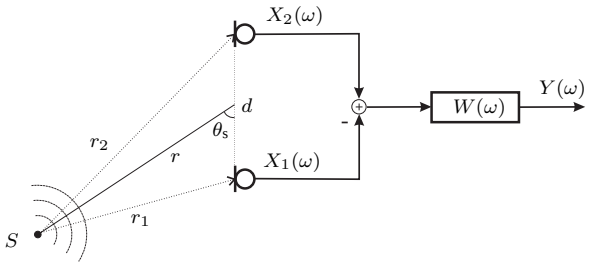

Fig. 2. Illustration of first-order ACTMA with a nearby source.

Assuming a spherical wave propagation model for the source S, the frequency-domain microphone signals can be modeled as [5]

$$\begin{aligned} X_i(\omega) &= S(\omega)H_i(\omega) + N_i(\omega) \\ &= S(\omega)\frac{e^{-j\omega r_i/c}}{r_i} + N_i(\omega), \quad i = 1, 2, \end{aligned} \tag{1}$$

where $S(\omega)$ is the desired speech signal, $H_i(\omega)$ is the transfer function from the desired source to the i-th microphone, $N_i(\omega)$ is the background noise and additive uncorrelated white noise, $\omega = 2\pi f$, and c is the speed of sound.

According to [5], the correction filter $W(\omega)$ is computed as the inverse of the nearfield response of the differential array to the source $S(\omega)$, which is given by

$$B(r, \theta_s; \omega) = \frac{e^{-j\omega r_1/c}}{r_1} - \frac{e^{-j\omega r_2/c}}{r_2}, \tag{2}$$

where r_1 and r_2 are a function of r and θ_s. The correction filter results in a nominally flat frequency response, thus ensuring the desired signal remains undistorted, without significantly degrading the noise canceling properties of CTMAs. Since the position of the mobile phone user is unkown, the correction filter is parameterized in practice by the estimated distance $\tilde{r}$ and angular orientation $\tilde{\theta}_s$ of the mobile phone user's mouth relative to the array axis. These parameters can be estimated as proposed in [5].

3. STEERED ACTMA

In mobile phone scenarios, the distance and angular orientation of the mobile phone user relative to the array varies significantly from user to user. As the null of the ACTMA is fixed at broadside, i.e., 90°, the correction filter may cause a significant amplification of the uncorrelated spatially white noise if the mobile phone user's angular position θ_s approaches 90°.

To avoid the problem stated above, we propose to use the steered ACTMA (SACTMA), which is depicted in Figure 3. The null of the SACTMA is constrained to an angular region in which the phone user is typically not found by introducing a delay $\tau(\theta_{\text{null}}) = d/c \cos\theta_{\text{null}}$ in the signal path. The null can be steered adaptively, e.g., by localizing the dominant interferer during periods of mobile phone user inactivity while also constraining the estimated DOA to a predefined angular region ,e.g., $120^\circ < \theta_{\text{null}} \leq 180^\circ$. In this paper, the null is fixed at an angle of $\theta_{\text{null}} = 180^\circ$.

For the SACTMA, we select the desired signal at microphone m_1, i.e., $\hat{S}(\omega) = S(\omega)\exp(-j\omega r_1/c)/r_1$, as our reference. We

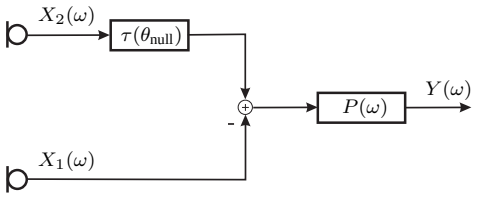

Fig. 3. First-order SACTMA.

therefore seek to estimate $\hat{S}(\omega)$ instead of $S(\omega)$. The inputs to the SACTMA may then be written as

$$X_1(\omega) = \hat{S}(\omega) + N_1(\omega), \tag{3}$$

and

$$\begin{aligned} X_2(\omega) &= \hat{S}(\omega)\frac{r_1}{r_2}e^{-j\omega(r_2-r_1)/c} + N_2(\omega) \\ &= \hat{S}(\omega)\sigma_{12}e^{-j\omega\tau_{12}} + N_2(\omega). \end{aligned} \tag{4}$$

In this case the correction filter $P(\omega)$ is obtained by computing the inverse of the response with respect to $\hat{S}(\omega)$, which is given by

$$\hat{B}(r, \theta_s; \omega) = 1 - \sigma_{12}e^{-j\omega(\tau_{12}+\tau(\theta_{\text{null}}))}, \tag{5}$$

instead of (2).

In order to compute the inverse of (5), we require an estimate of the distance ratio σ_{12} and the time-difference of arrival (TDOA) τ_{12} between the microphones. Similarly to [5], the distance ratio can be estimated by

$$\tilde{\sigma}_{12}(\kappa) = \lambda_1 \frac{\sum_\mu |X_2(\mu, \kappa)|}{\sum_\mu |X_1(\mu, \kappa)|} + (1 - \lambda_1)\tilde{\sigma}_{12}(\kappa - 1) \tag{6}$$

where λ_1 is a smoothing parameter. The discrete frequency bin and frame index are denoted by μ and κ, respectively. Note that a mismatch in the microphone gains results in a wrong estimate of the distance ratio. The TDOA τ_{12} can be estimated by any one of the various methods presented in the literature [9, 10]. Here, the TDOA is estimated by using the Generalized Cross Correlation (GCC) method [11].

4. ROBUST SACTMA

To ensure sufficient accuracy in the estimation of the parameters $\tilde{\sigma}_{12}$ and $\tilde{\tau}_{12}$, the estimation should only occur during periods when the mobile phone user is active. In addition, the impact of microphone mismatch on the SACTMA performance should be minimized.

In this section, we present a robust SACTMA algorithm, which seeks to overcome these challenges. Figure 4 depicts the block diagram of the proposed method. The source signals are captured by three microphones, i.e., m_1, m_2 and m_3, and the microphone signals are subsequently sampled and quantized, and then a filterbank is applied to obtain the frequency-domain signals $X_1(\mu, \kappa)$, $X_2(\mu, \kappa)$ and $X_3(\mu, \kappa)$.

4.1. Near/Far Activity Detector

In order to achieve sufficient parameter estimation accuracy and to perform online calibration, we require a method to distinguish between the activity of the mobile phone user and the background noise.

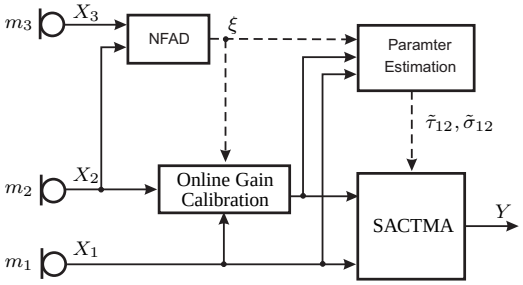

Fig. 4. General block diagram of proposed robust SACTMA.

In this section, we consider the near/far activity detector (NFAD) whose main goal is to distinguish between the presence of speech coming from the mobile phone user and the presence of background noise. This may be achieved by computing the NPLD between microphones m_2 and m_3 [4]

$$\Gamma(\mu,\kappa) = \left| \frac{\Phi_{x_2x_2}(\mu,\kappa) - \Phi_{x_3x_3}(\mu,\kappa)}{\Phi_{x_2x_2}(\mu,\kappa) + \Phi_{x_3x_3}(\mu,\kappa)} \right|, \tag{7}$$

where $\Phi_{x_ix_i}(\mu,\kappa) = \lambda_2 \left| X_i^2(\mu,\kappa) \right| + (1-\lambda_2)\Phi_{x_ix_i}(\mu,\kappa-1)$ are the power spectral densities (PSDs) estimates of $X_i(\mu,\kappa)$ and λ_2 is a smoothing parameter. It was shown in [4] that the NPLD contains information related to the proximity of a source with respect to the mobile phone. Note that $0 \leq \Gamma(\mu,\kappa) \leq 1$.

When only the background noise sources are active the power at the microphones is approximately equal and $\Gamma(\mu,\kappa)$ approaches zero. When the telephone user is active there is a significant difference in power at the microphones and therefore $\Gamma(\mu,\kappa)$ approaches unity. By applying a threshold to the NPLD, a decision ξ can be made on whether the telephone user is active or only the background noise sources are active. This information is subsequently used to control other modules as will be explained shortly.

The NPLD computation in (7) assumes that the gains of the microphones are matched. Unfortunately this is seldom the case in practice due to manufacturing tolerances. Actually gain mismatches introduce a bias into the NPLD computation. Assuming the microphone gains are constant over time, (7) becomes

$$\Gamma(\mu,\kappa) = \left| \frac{\Phi_{x_2x_2}(\mu,\kappa) - g_{32}(\mu)\Phi_{x_3x_3}(\mu,\kappa)}{\Phi_{x_2x_2}(\mu,\kappa) + g_{32}(\mu)\Phi_{x_3x_3}(\mu,\kappa)} \right|, \tag{8}$$

where $g_{32}(\mu) = g_3^2(\mu)/g_2^2(\mu)$ is the ratio of the gains of microphones m_3 and m_2, respectively. If the microphones capture background noise such that $\Phi_{x_2x_2} = \Phi_{x_3x_3}$ then (8) becomes

$$\Gamma_{\text{bg}}(\mu) = \frac{1 - g_{32}(\mu)}{1 + g_{32}(\mu)}. \tag{9}$$

For the algorithm proposed in [4], if the threshold $\gamma_{\text{min}} < \Gamma_{\text{bg}}(\mu)$ this would lead to infrequent updates of the power spectral density (PSD) estimate and therefore less noise reduction.

Figure 5 depicts an exemplary broadband NPLD computed from recorded signals. Note that for our purposes, the NPLD averaged over frequency, $\bar{\Gamma}(\kappa) = \langle \Gamma(\mu,\kappa) \rangle_\mu$, is sufficient for signal classification. The signals were recorded at a busy bus stop using a mock-up mobile phone whose microphones were located as depicted in Figure 1. The spacing of the microphones at the bottom was 5 mm. Although high NPLD values occur when the mobile phone user is active as expected, when only background noise is present the NPLD is shifted upwards due to microphone mismatch. This behavior was confirmed by other measurements in different acoustic environments and using different sets of microphones.

To improve robustness, we propose to track the minima of the broadband NPLD in order to compute an adaptive threshold, i.e., the threshold is set relative to the minimum NPLD. Tracking of the NPLD minima is performed similarly to the method proposed in [12]. The main idea is to find the minimum NPLD Γ_{min} within a predefined number of consecutive frames. The adaptive threshold, depicted in Figure 5, is given by $\gamma_{\text{amin}}(\kappa) = \Gamma_{\text{min}}(\kappa) + \gamma_{\text{min}}$, where γ_{min} is a fixed threshold.

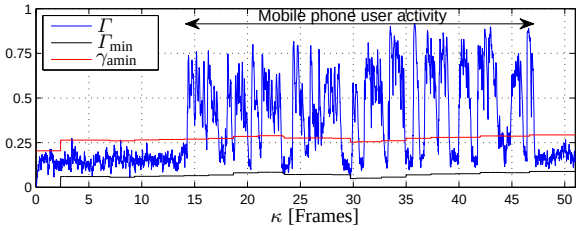

Fig. 5. Exemplary NPLD, minimum NPLD, and adaptive threshold.

4.2. Online Gain Calibration

It is well known that microphone mismatch and position errors lead to a significant degradation in the performance of ACTMAs. In [5] the authors suggested performing an offline calibration in order to reduce the microphone mismatch. Although effective, this procedure is not feasible for mass produced mobile phones.

In this contribution, we propose online gain calibration because experiments showed that the performance degradation due to gain mismatch is significantly greater than due to phase mismatch. Although gain mismatches are frequency-dependent in practice, a frequency-independent (broadband) calibration gain is used here. The gain calibration module computes broadband gains that compensate for microphone gain mismatches, i.e., typically less than ± 3 dB, between microphones m_1 and m_2. The gain calibration works on the assumption that if only the background noise is active, the power of the signals at microphone m_1 and m_2 should be the same. This is a reasonable assumption as the microphones are very close to each other. The broadband calibration gains are computed as

$$g_{12}(\kappa) = \lambda_3 \frac{\bar{\Phi}_{x_1x_1}(\kappa)}{\bar{\Phi}_{x_2x_2}(\kappa)} + (1-\lambda_3) g_{12}(\kappa-1) \tag{10}$$

if $\bar{\Gamma}(\kappa) \leq \gamma_{\text{amin}}(\kappa)$, where $\bar{\Phi}_{x_ix_i}(\kappa) = \sum_\mu \Phi_{x_ix_i}(\mu,\kappa)$, λ_3 is a smoothing parameter, and $\gamma_{\text{amin}} = 0.2$ was chosen empirically.

4.3. Robust Parameter Estimation

The accurate estimation of the distance ratio σ_{12} and the TDOA τ_{12}, which are used in the computation of the correction filter as was explained in Section 3, is important as this minimizes the distortion of the speech from the mobile phone user. If the parameter estimation were to be performed continuously, this would lead to spurious estimates and a degradation in performance. Additionally, microphone gain mismatch leads to erroneous distance ratio estimates.

Therefore, the parameter estimation module estimates the distance ratio $\tilde{\sigma}_{12}$ and the TDOA $\tilde{\tau}_{12}$ between $X_1(\mu,\kappa)$ and

$\hat{X}_2(\mu,\kappa) = \sqrt{g_{12}} X_2(\mu,\kappa)$ only when speech activity of the mobile phone user is detected by the NFAD, i.e., if $\bar{\Gamma}(\kappa) \geq \gamma_{\text{amin}}(\kappa) + \delta$, where the value $\delta = 0.4$ was chosen empirically.

5. PERFORMANCE EVALUATION

First we compare the performance of the ACTMA and SACTMA algorithms with respect to mobile phone usere's DOA θ_s. The performance is evaluated using the signal-to-interference-plus-noise ratio (SINR) gain, which is defined as the ratio of the segmental SINR at the algorithm's output w.r.t. the segmental SINR at the reference microphone m_1. The microphone signals were obtained by convolving audio files with room impulse responses that were generated by the image method [13] for a room with dimensions 5x5x2.5 m and a reverberation time of 350 ms. A sampling frequency of 32 kHz and microphone spacing of 5 mm were chosen. The desired source was placed at a distance of 7.5 cm from the center of the array. An interferer was placed at a distance of 2 m at an angle of $60°$. Here, we assume that the DOA and distance of the desired source is known. Figure 6 depicts the gains of the ACTMA and SACTMA with respect to θ_s. Clearly, the gain decreases for both methods as the desired source moves towards broadside but the SACTMA has superior performance.

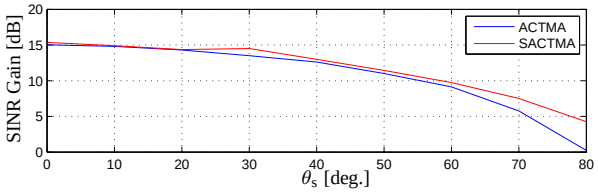

Fig. 6. SINR gain of ACTMA and SACTMA with respect to angular orientation of mobile phone user.

Now we investigate, by way of examples, the effect of broadband gain calibration on the algorithmic performance. For this, the phase and magnitude responses of forty five EPCOS C914G MEMS microphones were used. The response for microphone m_1 was computed from mean magnitude and phase responses, i.e., $H_1(\mu) = \bar{g}_\text{r}(\mu)\exp(j\omega_\mu \bar{\phi}_\text{r}(\mu))$. The response of the other microphones $i = 2, 3$ were obtained as the realization of a Monte Carlo experiment with Gaussian distributions for amplitude and phase:

$$H_{i,q}(\mu) = \left(\bar{g}_\text{r}(\mu) + \frac{\sigma_\text{m}(\mu)}{\sigma_\text{m}(\mu_0)}\Delta g_{i,q}\right) e^{-j\omega_\mu\left(\bar{\phi}_\text{r}(\mu) + \frac{\sigma_\text{p}(\mu)}{\sigma_\text{p}(\mu_0)}\Delta\phi_{i,q}\right)} \tag{11}$$

where q is one of Q realizations, $\sigma_\text{m}(\mu)$ is the measured standard deviation for bin μ, and $\sigma_\text{m}(\mu_0)$ is the measured standard deviation for an arbitrary reference bin μ_0 (here the bin corresponds to 1 kHz). $\Delta g_{i,q}$ and $\Delta\phi_{i,q}$ are the zero-mean Gaussian distributed magnitude and phase errors with a variance of σ_m^2 and σ_p^2, respectively.

Figure 7 illustrates the improvement in SINR obtained from the online gain calibration compared to the uncalibrated case. The results were obtained by averaging twenty realizations for each chosen variance pair $(\sigma_\text{m}^2, \sigma_\text{p}^2)$. It is clear that the application of gain calibration improves the performance of the algorithm significantly, up to almost 3 dB. It should be noted that for very small gain deviations of less than 0.01 dB , the broadband gain calibration leads to minimal performance degradation. Further improvement might be achieved by performing frequency dependent gain calibration, which is a topic of future research.

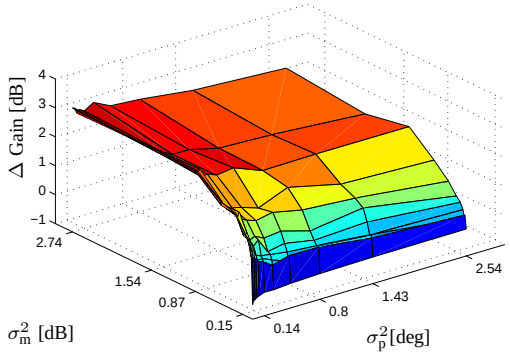

Fig. 7. Robust SACTMA SINR gain improvement due to broadband gain calibration.

Finally we evaluate the performance of the robust SACTMA for real recordings. Figure 8 depicts the input PSD of the signal recorded by microphone m_1 and the output PSD the robust SACTMA for real recordings performed at a busy bus stop (see Section 4.1 for further details). Note that the DOA and distance of the desired source to the array are unknown and have to be estimated in this case. It is clear that robust SACTMA achieves significant background noise reduction. The residual noise at low frequencies is predominantly spatially white noise. This residual noise can be reduced significantly by applying single-channel noise reduction [3] as a postprocessing step to further reduce residual noise.

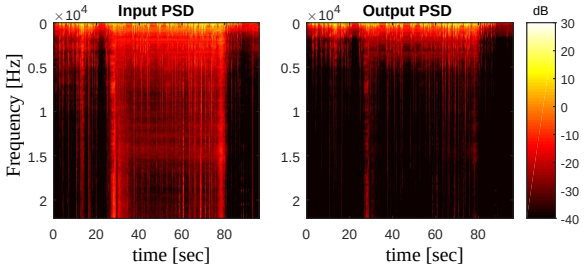

Fig. 8. Robust SACTMA input and output PSDs.

6. ACKNOWLEDGEMENTS

The authors would like to thank EPCOS AG and the Munich University of Applied Sciences for providing the magnitude and phase response measurements for the EPCOS C914G MEMS microphones.

7. CONCLUSIONS

In this paper we have proposed a method that improves the robustness of the ACTMA algorithm by performing robust parameter estimation and online calibration. We also showed that it is necessary to take the microphone gain mismatch into account when using the NPLD for signal classification. Experimental results confirmed the applicability of robust SACTMA for performing noise reduction in mobile phones.

8. REFERENCES

[1] L. Watts, "Advanced noise reduction for mobile telephony," in *IEEE Computer Magazine*, August 2008, vol. 41, p. 9092.

[2] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Processing*, pp. 504–512, 2001.

[3] T. Gerkmann and R.C. Hendriks, "Unbiased-MMSE based noise power estimation with low complexity and low tracking delay," in *IEEE Transactions on Audio, Speech & Language Processing*, March 2012, vol. 20, pp. 1383–1393.

[4] M. Jeub, C. Herglotz, C.M. Nelke, C. Beaugeant, and P. Vary, "Noise reduction for dual-microphone mobile phones exploiting power level differences," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, March 2012, pp. 1693–1696.

[5] H. Teutsch and G. Elko, "An adaptive close-talking microphone array," in *Proc. IEEE WASPAA*, October 2001, pp. 21–24.

[6] J. Benesty and J. Chen, Eds., *Study and Design of Differential Microphone Arrays*, Springer-Verlag, Berlin, Germany, 2013.

[7] W.W. Hansen and J.R. Woodyard, "A new principle in directional antenna design," in *Proc. IRE*, March 1938, vol. 26, pp. 333–345.

[8] S.A. Schelkunoff, "A mathematical theory of linear arrays," in *Bell Syst. Tech. J.*, January 1943, vol. 2, pp. 80–107.

[9] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 3, pp. 288–292, May 1997.

[10] M. Souden, J. Benesty, and S. Affes, "Broadband source localization from an eigenanalysis perspective," *IEEE Trans. Speech and Audio Processing*, vol. 18, no. 6, pp. 1575–1587, August 2010.

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

[12] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. Euro. Signal Processing Conf. (EUSIPCO)*, October 1994, pp. 1182–1185.

[13] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.