PITCH AND TDOA-BASED LOCALIZATION OF ACOUSTIC SOURCES WITH DISTRIBUTED ARRAYS

Martin Weiss Hansen, Jesper Rindom Jensen and Mads Græsbøll Christensen

Audio Analysis Lab, AD:MT, Aalborg University, Denmark
{mwh,jrj,mgc}@create.aau.dk

ABSTRACT

In this paper, a method for acoustic source localization using distributed microphone arrays based on time-differences of arrival (TDOAs) is presented. The TDOAs are used to estimate the location of an acoustic source using a recently proposed method, based on a 4D parameter space defined by the 3D location of the source, and the TDOAs. The performance of the proposed method for acoustic source localization is compared to the performance of a method based on generalized cross-correlation with phase transform (GCC-PHAT) using synthetic and speech signals with varying source position. Results show a decrease in the error of the estimated position when the proposed method is used.

Index Terms— Acoustic source localization, distributed microphone arrays, TDOA estimation, pitch estimation.

1. INTRODUCTION

Microphone arrays are found in many modern-day devices such as hearings aids, smartphones, smart TVs, laptops, game consoles and robots. These microphone arrays facilitate beamforming [1] and direction of arrival (DOA) estimation, and can be used for, e.g., video teleconferencing [2]. When multiple microphones are available, it is possible to perform acoustic source localization, where the range and DOA of the acoustic source are estimated jointly.

For the problem of source localization, several classes of methods exist, like multilateration [3], TDOA-based source localization [4], maximum likelihood source localization [5] and energy-based source localization [6]. Some of the earliest work is for SONAR applications, an example being [7]. In [4], a method based on range differences (RDs) is used to estimate the position of an acoustic source using a single array. In [8–10], source localization strategies are reviewed and categorized, and time-delay estimation (TDE) and source-detection is considered. More recent work takes advantage of distributed microphone arrays. In [11], distributed microphone arrays are used to localize an acoustic source, by using hyperbolic constraints. TDOAs are estimated using GCC-PHAT [12]. A class of geometric methods for acoustic source localization based on TDOAs is described in [13]. In the paper, two categories of solutions to the acoustic source localization problem using TDOAs are described, based on maximum likelihood (ML) and least squares (LS) criteria. If multiple independent microphone arrays are distributed in an environment, these are likely not to be synchronized. This aspect can be of great importance when estimating TDOAs, and in [14] a solution to the problem of synchronization between independent microphone arrays is presented.

It appears from the above that many localization methods are based on TDOA estimation, which is often done using GCC-PHAT. Examples can be found in [15,16]. Another possible solution is to jointly estimate the TDOAs and the pitch, since many audio signals have a harmonic structure that could be exploited. The latter approach can be used to get statistically efficient estimates, which is generally not possible with correlation based methods such as GCC-PHAT [17]. This is the approach taken herein, although we assume the pitches to be known, since pitch estimation is not the main topic of this paper.

In this paper, we propose an acoustic source localization scheme for distributed, unsynchronized uniform linear arrays (ULAs) of microphones, that uses the cone-based localization method of [13, 14]. The method is based on ML TDOA estimation for each channel. We propose to estimate the TDOAs of the sensor signals individually for each array using a maximum likelihood approach, inspired by [17], instead of the often used GCC-PHAT method. In this paper, the locations of the microphone arrays and the sensors are assumed to be known. If the locations of the microphone arrays are unknown, they can be estimated using a method for automatic microphone localization, like the one in [18].

The paper is organised as follows. In Section 2, the signal model is introduced. In Section 3, the proposed method of using ML TDOA estimation in cone-based localization for acoustic source localization with distributed microphone arrays is described. Section 4 presents the experimental setup and results, and the work is concluded in Section 5.

This work was supported in part by the Villum Foundation and the Danish Council for Independent Research, grant ID: DFF 1337-00084.

2. SIGNAL MODEL

We will now introduce the signal model for each microphone array. Consider a single, quasi-periodic source, such as speech, being sampled spatially by a microphone array, consisting of N_s sensors at time n_t. For simplicity, we will assume that the microphones are arranged as a uniform linear array (ULA). If the input signal is denoted $x(n_t)$, the output at sensor n_s in each array is

$$y_{n_s}(n_t) = x_{n_s}(n_t) + w_{n_s}(n_t), \tag{1}$$

with $x_{n_s}(n_t) = s(n_t - f_s\tau_{n_s})$, where τ_{n_s} is the TDOA between the reference sensor and sensor n_s, f_s is the sampling frequency, $s(n_t - f_s\tau_{n_s})$ is the delayed signal and $w_{n_s}(n_t)$ is the noise recorded by sensor n_s. It is assumed that the distance between sensors is small, such that there is no attenuation between channels. The signal is assumed to be harmonic, thus it can be modelled as a sum of complex sinusoids

$$s(n_t) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_0 n_t}, \tag{2}$$

where L is the model order, $\alpha_l = A_l e^{j\phi_l}$, where $A_l > 0$ is the real amplitude, and ϕ_l is the phase of the lth harmonic. It should be noted that the model (2) can be used for all periodic signals by careful selection of L and ω_0. The model can be applied to real data by using the Hilbert transform. Using (2), the signal at sensor n_s is

$$s(n_t - f_s\tau_{n_s}) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_0 n_t} e^{-jl\omega_s n_s}, \tag{3}$$

where $\omega_s = \omega_0 f_s \tau_1$ is the spatial frequency. It is assumed that the ULA is placed in the far-field of the source, and that the recording environment is anechoic. A spatio-temporal vector signal model of the signal in (1) can then be defined as

$$\mathbf{y} = \mathbf{x} + \mathbf{w} = \mathbf{Z}\boldsymbol{\alpha} + \mathbf{w}, \tag{4}$$

where $\mathbf{y} = [y_0(n_t) \quad \cdots \quad y_{N_s-1}(n_t) \quad \cdots \quad y_0(n_0 - N_t + 1) \quad \cdots \quad y_{N_s-1}(n_t - N_t + 1)]^T$, with similar definitions for $\mathbf{x}$ and $\mathbf{w}$, and

$$\mathbf{Z} = [\mathbf{z}(\omega_0, \omega_s) \cdots \mathbf{z}(L\omega_0, L\omega_s)], \tag{5}$$

$$\mathbf{z}(l\omega_0, l\omega_s) = \mathbf{z}_t(l\omega_0) \otimes \mathbf{z}_s(l\omega_s), \tag{6}$$

$$\boldsymbol{\alpha} = [\alpha_1 e^{j\omega_0 n_t} \cdots \alpha_L e^{jL\omega_0 n_t}]^T, \tag{7}$$

$$\mathbf{z}_s(\omega_s) = \begin{bmatrix} 1 & e^{-j\omega_s} & \cdots & e^{-j(N_s-1)\omega_s} \end{bmatrix}^T, \tag{8}$$

$$\mathbf{z}_t(\omega_0) = \begin{bmatrix} 1 & e^{-j\omega_0} & \cdots & e^{-j(N_t-1)\omega_0} \end{bmatrix}^T, \tag{9}$$

where $\otimes$ denotes the Kronecker product. With the signal model in place, we can proceed to the estimation of the TDOAs for each array and the proposed localization scheme that takes into account the TDOAs for multiple arrays.

3. PROPOSED METHOD

The proposed method for source localization is based on the cone-based localization method of [13, 14]. In this paper, however, a maximum likelihood (ML) TDOA estimator, inspired by [17], is used to estimate TDOAs instead of the standard approach of using GCC-PHAT.

In the case of white Gaussian noise, with equal variance on all channels, the ML estimator for TDOA estimation is asymptotically equivalent to the non-linear least squares (NLS) method [17]. The NLS TDOA estimates for each array are found by solving

$$\hat{\tau} = \underset{\boldsymbol{\alpha}, \tau \in \mathrm{T}}{\arg\min} \|\mathbf{y} - \mathbf{Z}\boldsymbol{\alpha}\|_2^2, \tag{10}$$

where T are the possible TDOAs. To estimate the unknown amplitudes $\boldsymbol{\alpha}$, (10) is maximized with respect to the parameter in question, resulting in the following estimates

$$\hat{\boldsymbol{\alpha}} = (\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H\mathbf{y}. \tag{11}$$

If (11) is inserted into (10), we find

$$\hat{\tau} = \underset{\tau \in \mathrm{T}}{\arg\max}\, \mathbf{y}^H\mathbf{Z}(\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H\mathbf{y}. \tag{12}$$

Using the above estimator, we can estimate the TDOA τ_{n_s} of a quasi-periodic signal, with known pitch and model order, sampled by each ULA. Because of the nonlinear nature of the cost function (12), a grid search is used to evaluate the cost function. In this paper, the pitch ω_0 of the signal is assumed to be known. It can easily be estimated in practice, e.g., using the fast, multichannel, FFT-based method in [19].

Equipped with the TDOA estimates $\hat{\tau}_{n_s}$ for each sensor in each of the arrays, we proceed by using the source localization method [13]. In this localization method, an extended coordinate system is formed by adding a range difference coordinate to the coordinates of the source. Each point $\mathbf{p} = [x, y, s]^T$ is hereby mapped onto the 4D space-range $[\mathbf{p}^T, w]^T$, where w is the range coordinate. Consider a ULA with N_s sensors, placed at $\mathbf{m}_{n_s} = [x_{n_s}, y_{n_s}, z_{n_s}]^T$, where $n_s = 0, \ldots, N_s - 1$. The range difference between the source and the n_sth microphone and between the source and the reference microphone is

$$w_{n_s} = c\tau_{n_s,0} = \|\mathbf{p}_s - \mathbf{m}_{n_s}\| - \|\mathbf{p}_s - \mathbf{m}_0\|, \tag{13}$$

where $\mathbf{m}_0$ is the reference sensor, $\mathbf{p}_s$ is the position of the source, and τ_{n_s} is found using (12). The microphones $\mathbf{m}_{n_s}$ can be represented by cones with apex $[\mathbf{m}_{n_s}, w_{n_s}]^T$. With noiseless measurements, $\mathbf{p}_s$ should fall in the intersection of all such cones (see [13] for further details). Because of noisy measurements, the source location is found as the point $\widehat{\mathbf{p}}_s$ with minimum distance to the surface of the cones, i.e.,

$$\widehat{\mathbf{p}}_s = \underset{\mathbf{p}_s \in \mathbf{P}_s}{\arg\min} \|\mathbf{e}(\mathbf{p}_s)\|^2, \tag{14}$$

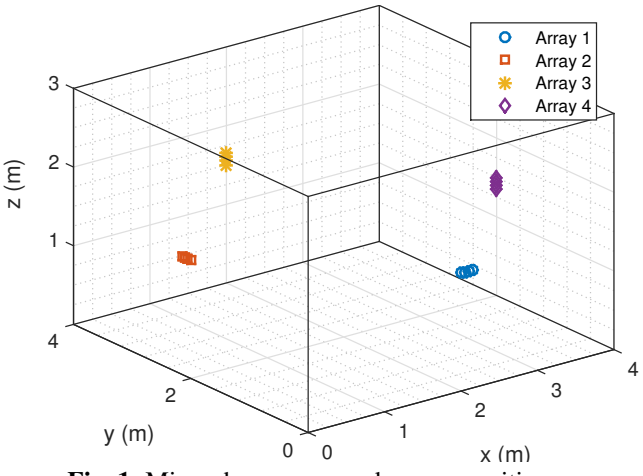

Fig. 1. Microphone array and source positions.

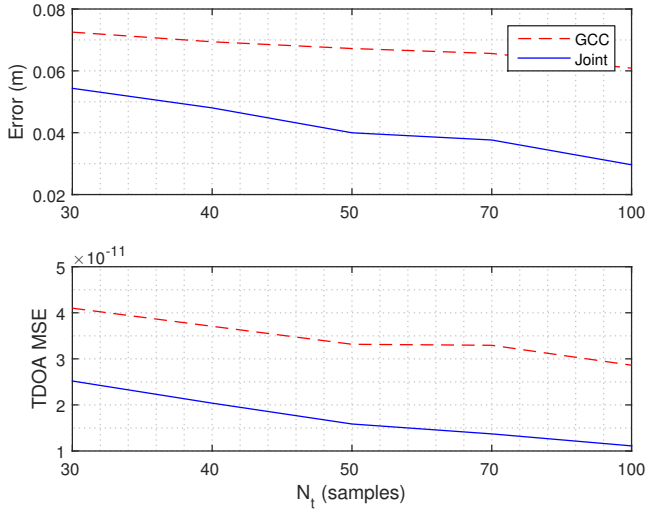

Fig. 2. Average of the magnitude of localization error in m (top) and TDOA MSE (bottom) for varying segment lengths, using a synthetic signal.

where $\mathbf{P}_s$ is a set of candidate source positions,

$$\mathbf{e}(\mathbf{p}_s) = [e_0(\mathbf{p}_s) \cdots e_{N_s-1}(\mathbf{p}_s)]^T, \tag{15}$$

and $e_{n_s}(\mathbf{p}_s) = (x_s - x_{n_s})^2 + (y_s - y_{n_s})^2 + (z_s - z_{n_s})^2 - (w_s - w_{n_s})^2$. If the microphones in the arrays are not synchronized, the z-axis does not correspond to the reference microphone. To address synchronization in the case of N_s distributed arrays, a difference between the distance from the source to the reference and to the local reference of the n_sth array is defined as [14]

$$\Delta z^{(n_s)} = \sqrt{\Delta_{x_{n_s}}^2 + \Delta_{y_{n_s}}^2 + \Delta_{z_{n_s}}^2} - \sqrt{\Delta_{x_0}^2 + \Delta_{y_0}^2 + \Delta_{z_0}^2}, \tag{16}$$

where $\Delta_{x_{n_s}} = x_{0,n_s} - x_s$, x_{0,n_s} is the position of the reference sensor in the n_sth array. By adding the displacements (16) to the cone errors in (14), the range difference estimates refer to a single global reference microphone.

It should be noted that array placement is an important factor to consider. In this paper, since 3D localization is considered, the environment must contain arrays that estimate the position in three dimensions, in order to estimate the position of an acoustic source in 3D. For further details, see [13].

4. EXPERIMENTAL RESULTS

The performance of the proposed method of ML TDOA estimation, as described in Section 2, used for source location estimation, is compared to GCC-PHAT [12, 20]. Both methods have been evaluated using an anechoic synthetic harmonic signal, consisting of 10 harmonic complex sinusoids, with a sampling rate of $f_s = 8$ kHz. The speed of sound is assumed to be $c = 340$ m/s. The acoustic environment is set up using the Signal Generator for MATLAB [21], which is based on the image method [22]. The room dimensions are 4 by 4 by 3 m. Four microphone arrays each consisting of four microphones are used. The arrays are placed with their reference microphones at the middle of each side of the room, along the walls. The microphone spacing is 5 cm. Figure 1 shows the placement of the microphones. After generating a multichannel signal with the above-mentioned setup, a number of channels of diffuse white Gaussian noise, corresponding to the total number of microphones, is added to the signals received at the microphones, resulting in an SNR at the microphones, SNR_m. White Gaussian noise is added to the source signal, resulting in a varying SNR at the source position, SNR_s.

As mentioned earlier, TDOA estimation is performed using an ML TDOA estimation technique. The signals are processed individually for each array, since we are considering a distributed network of microphone arrays. The maximum possible TDOA corresponds to the distance between the sensors in the array. Because of this, the TDOAs are estimated using a grid ranging from -0.1875 to 0.1875 ms, corresponding to a range of -1.5 to 1.5 samples. The search grid spacing is 0.01 samples. For GCC-PHAT the same TDOA grid was used. Furthermore, an FFT length of 512 samples was used, and in this method we integrate over frequencies in the range $f = [300, 4000]$ Hz [23].

Two experiments were conducted in order to assess the performance of the proposed method; one where the segment length N_t was varied, and one where SNR_m was varied. The data is obtained by conducting 500 Monte-Carlo simulations for each data point. In each simulation, the source position is randomly chosen, $x \in [1.0, 3.0]$, $y \in [1.0, 3.0]$ and $z \in [1.0, 2.0]$ with a search grid spacing of 1 cm in all directions. Furthermore, the fundamental frequency is sampled from the interval $f_0 \in [300, 400]$ Hz, and the phase of each of the harmonics is randomized. Figure 2 shows the magnitude of the

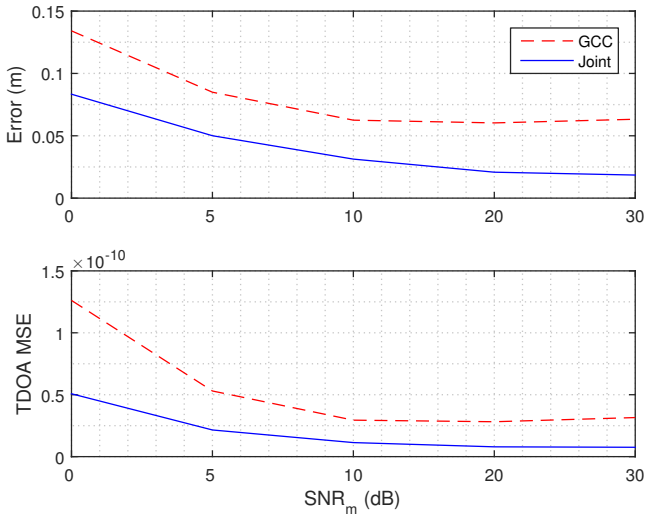

Fig. 3. Average of the magnitude of localization error in m (top) and TDOA MSE (bottom) for varying SNRs, using a synthetic signal.

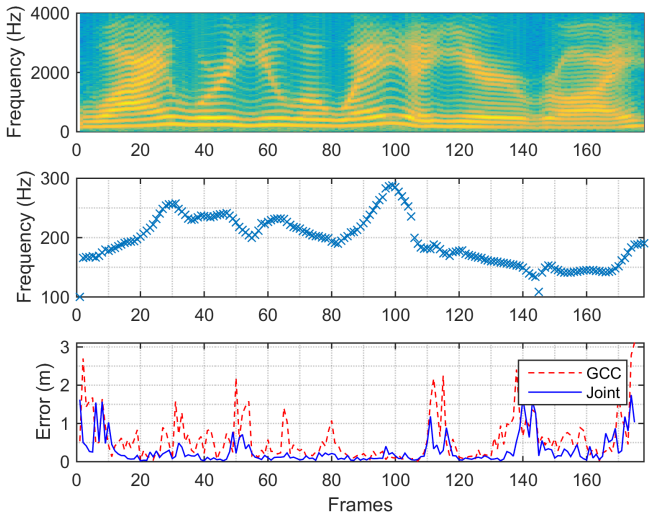

Fig. 4. Spectrogram of a moving speech signal (top), estimated f_0 of the signal (middle), and magnitude of localization error in m (bottom).

localization error in meters, and the TDOA mean squared error (MSE), for segment lengths varying from 30 to 100 samples. The SNR of the signal emitted from the position of the acoustic source is $\text{SNR}_s = 20$ dB. The SNR of the signals received at the microphones is $\text{SNR}_m = 10$ dB. The model order is $L = 10$. The figure shows a decrease in the localization error, for both methods, when the segment length is increased. Furthermore, the error when using GCC-PHAT is approximately twice as large compared to when the ML TDOA estimation method is used. Figure 3 shows the magnitude of the localization error in meters, and the TDOA MSE, for SNR_m varying from 0 to 30 dB. For this experiment, the segment length is $N_t = 100$ samples, $\text{SNR}_s = 20$ dB, and the model order is $L = 10$. The figure shows a decrease in the magnitude of the localization error in m, for both methods, when SNR_m is increased. Furthermore, the error when using GCC-PHAT is approximately twice as large compared to when ML TDOA estimation is used. In both experiments, the results show that using ML TDOA estimates results in smaller errors than GCC-PHAT.

Furthermore, in order to qualitatively assess the performance of the proposed method using a real signal, an experiment was conducted using a speech signal ("Why were you away a year, Roy?"). The fundamental frequency of the signal was estimated using the joint ANLS method, from the same toolbox. The search interval for the fundamental frequency was $f_0 = [100, 500]$ Hz, and the FFT length was 16384 samples. Note that the signal was down-sampled from 44.1 kHz to 8 kHz. The speech signal was set to move in a straight line from $[2, 2, 1.5]$ m to $[3, 3, 1.75]$ m. The signal was processed one frame at the time, using non-overlapping frames of length $N_t = 100$ samples. The reflection order was set to 0. In this experiment, $\text{SNR}_s = 20$ dB, and $\text{SNR}_m = 10$ dB.

Figure 4 shows the spectrogram of the speech signal, the estimated fundamental frequency for each frame and the localization error in meters for each frame. The mean error for the ML TDOA method is 0.30 m, and 0.59 m for the GCC-PHAT method. This reduction in localization error is consistent with the results using synthetic signals.

5. DISCUSSION

In this paper, the problem of acoustic source localization based on TDOA estimation is considered. In particular, the method of using ML TDOA estimation, inspired by [17], is compared to using GCC-PHAT [12]. The considered scenario consists of multiple distributed microphone arrays. The performance of the source localization method when using ML TDOA estimation is compared to the performance when using GCC-PHAT for TDOA estimation. By using ML TDOA estimation, the accuracy of the acoustic source localization is increased, when compared to using GCC-PHAT, using both synthetic and real signals. This is expected, since the TDOA estimator is the maximum likelihood estimator, when the noise is white Gaussian, with the same variance on each channel, the environment is anechoic and the source is in the far-field. The results are of particular interest for localization of moving sources, since the ML TDOA estimation method results in small errors even at low segment lengths, which are required for moving sources, and low SNRs. Furthermore, an accurate estimate of the position of an acoustic source is important for enhancement purposes.

6. REFERENCES

[1] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[2] B. Kapralos, M. R. M. Jenkin, and E. Milios, "Audio-visual localization of multiple speakers in a video teleconferencing setting," *Int. J. of Imaging Systems and Technology*, vol. 13, no. 1, pp. 95–105, 2003.

[3] T. E. Tuncer and B. Friedlander, *Classical and Modern Direction-of-Arrival Estimation*, Academic Press, 2009.

[4] J. O. Smith and J. S. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 12, pp. 1661–1669, 1987.

[5] J. C. Chen, R. E. Hudson, and K. Yao, "Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field," *IEEE Trans. Signal Process.*, vol. 50, no. 8, pp. 1843–1854, August 2002.

[6] D. Li, K. D. Wong, Y. H. Hu, and A. M. Sayeed, "Detection, classification, and tracking of targets," *IEEE Signal Process. Mag.*, vol. 19, no. 2, pp. 17–29, March 2002.

[7] W. C. Knight, Roger G. Pridham, and S. M. Kay, "Digital signal processing for SONAR," *Proc. IEEE*, vol. 69, no. 11, pp. 1451–1506, 1981.

[8] M. S. Brandstein and H. F. Silverman, "A practical methodology for speech source localization with microphone arrays," *Comput. Speech Language*, vol. 11, no. 2, pp. 91 – 126, 1997.

[9] Y. Huang, J. Benesty, and G. W. Elko, "Source localization," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, pp. 229–253. Springer US, 2004.

[10] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, no. 1, pp. 1–19, 2006.

[11] A. Canclini, E. Antonacci, A. Sarti, and S. Tubaro, "Acoustic source localization with distributed asynchronous microphone networks," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 2, pp. 439–443, 2013.

[12] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[13] P. Bestagini, M. Compagnoni, F. Antonacci, A. Sarti, and S. Tubaro, "TDOA-based acoustic source localization in the space-range reference frame," *Multidim. Syst. Sign. Process.*, pp. 1–23, March 2013.

[14] M. Compagnoni et al., "Localization of acoustic sources through the fitting of propagation cones using multiple independent arrays," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 7, pp. 1964–1975, 2012.

[15] D. Salvati, S. Canazza, and A. Rodà, "A sound localization based interface for real-time control of audio processing," *Proc. Int. Conf. Digital Audio Effects*, pp. 177–184, 2011.

[16] P. Annibale and R. Rabenstein, "Closed-form estimation of the speed of propagating waves from time measurements," *Multidim. Syst. Sign. Process.*, pp. 1–18, 2013.

[17] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Nonlinear least squares methods for joint DOA and pitch estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 5, pp. 923–933, 2013.

[18] N. D. Gaubitch, W. B. Kleijn, and R. Heusdens, "Auto-localization in ad-hoc microphone arrays," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 106–110, 2013.

[19] M. G. Christensen, "Multi-channel maximum likelihood pitch estimation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 409–412, 2012.

[20] D. Hertz and M. Azaria, "Time delay estimation between two phase shifted signals via generalized cross-correlation methods," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 8, no. 2, pp. 235 – 257, 1985.

[21] E. A. P. Habets, "Signal generator for MATLAB," Tech. Rep., Technische Universiteit Eindhoven, 2011.

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

[23] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, pp. 157–180. Springer, 2001.