Robust acoustic source localization based on modal beamforming and time–frequency processing using circular microphone arrays

Ana M. Torres[a)]
I.E.E.A.C. Department, Universidad Castilla-La Mancha, 16071, Cuenca, Spain

Maximo Cobos
Computer Science Department, Universitat de València, 46100, Burjassot, Valencia, Spain

Basilio Pueo
Communication and Social Psychology Department, Universidad de Alicante, 03690, Alicante, Spain

Jose J. Lopez
Communications Department, Universitat Politècnica de València, 46022, Valencia, Spain

(Received 3 April 2012; revised 5 July 2012; accepted 11 July 2012)

Uniform circular array processing has been shown to be a very useful tool for broadband acoustic source localization over 360°. Specifically, beamforming methods based on circular harmonics have attracted a lot of research attention in the last several years, as modal array signal processing is a very active research topic. On the other hand, due to the sparsity properties of speech, source localization methods in the time–frequency (*T–F*) domain have also demonstrated their capability to locate several simultaneous sources with high accuracy. In this paper, a localization framework based on circular harmonics beamforming and *T–F* processing that provides accurate localization performance under very adverse acoustic conditions is presented. Modal processing and sparsity-based localization are jointly addressed to estimate the direction-of-arrival of multiple concurrent speech sources. Experiments in real and simulated environments with different microphone setups are discussed, showing the validity of the proposed approach and comparing its performance with other state-of-the-art methods. © *2012 Acoustical Society of America*.

[http://dx.doi.org/10.1121/1.4740503]

PACS number(s): 43.60.Fg [SAF] Pages: 1511–1520

I. INTRODUCTION

Acoustic source localization using microphone arrays is one of the most active research topics in multichannel signal processing. Microphone arrays have applications in many different areas such as immersive environments, human–computer interfaces, teleconferencing, or robot artificial audition.[1–5] However, broadband source localization under high noise and reverberation still remains a very challenging task. In recent years, new algorithms have been proposed to deal with this problem, making use of different array geometries and localization strategies.[6–8] In this context, modal processing using uniform circular arrays (UCA) has received increasing attention.[9] Methods based on circular harmonics beamforming (CHB) have shown to provide better localization performance than classical beamforming approaches. In fact, CHB belongs to a more recent class of methods often referred to as eigenbeamforming.[10–12] Tiana-Roig *et al.*[13] showed that CHB achieves better resolution and sidelobe properties than delay-and-sum beamforming (DSB) by selectively processing a different number of phase modes or spatial Fourier terms.

Besides beamforming-based localization methods, a number of algorithms working in the time–frequency (*T–F*) domain have also been recently developed.[14–16] Due to the sparse properties of speech in this domain, these methods are capable of localizing multiple active sound sources in real environments, even in those cases when the number of sources exceeds the number of microphones, i.e., underdetermined cases. To this end, inter-channel phase differences between microphone signals are analyzed to estimate the direction-of-arrival (DOA) of the dominant sound source at each *T–F* bin. The directions of the sources are finally estimated by fitting a specific model to the observed distribution of DOA estimates.

In this paper, we present a source localization method based on the combination of CHB and *T–F* processing. CHB is applied over each *T–F* point to estimate the DOA of the most dominant source by using a regularization-based approach. In contrast to conventional sparsity-based localization methods, the DOA estimates at each *T–F* bin are obtained by taking the direction of maximum CHB output power. This processing results in accurate DOA estimates under high reverberation and low signal-to-noise ratio (SNR) situations. Experiments considering different number of sources, microphones, reverberation degrees, and noise conditions are discussed. Moreover, the method is compared to other baseline localization techniques developed for UCA processing. The results show that the proposed approach is capable of localizing multiple sound sources in very reverberant and noisy environments with high accuracy.

[a)] Author to whom correspondence should be addressed. Electronic mail: ana.torres@uclm.es

The paper is structured as follows. Section II describes the theoretical background of CHB. Section III presents our proposed approach to CHB-based DOA estimation in the *T–F* domain. The description of the experiments and the discussion of the results are given in Secs. IV and V, respectively. Finally, the conclusions of this paper are summarized in Sec. VI.

II. CIRCULAR HARMONICS BEAMFORMING

A. Array geometry

Consider an UCA having M elements at equidistant locations on a circle of radius r lying on the xy plane, as shown in Fig. 1, where the center of the array is located at the origin of coordinates. The location vector of each element in Cartesian coordinates is given by

$$\mathbf{p}_m = [r\cos(\theta_m), r\sin(\theta_m), 0]^T, \quad m = 0, 1, ..., M-1, \tag{1}$$

where $(\cdot)^T$ denotes transposition and the azimuth angle of each element is $\theta_m = m(2\pi/M)$. The inter-element distance can be calculated as

$$d = 2r\sin\left(\frac{\pi}{M}\right). \tag{2}$$

The above distance determines the spatial aliasing frequency, which is given by

$$f_{\mathrm{al}} = \frac{c}{2d}. \tag{3}$$

Assuming signals coming from the median plane ($\phi_i = \pi/2$), the steering vector for the UCA depends on the azimuth angle as follows:

$$a(kr,\theta_i) = [e^{jkr\cos(\theta_i-\theta_0)}, ..., e^{jkr\cos(\theta_i-\theta_{M-1})}]^T, \tag{4}$$

where $k = 2\pi f/c$, f being the frequency and c the speed of sound ($c \approx 342$ m/s).

B. Circular apertures

1. Continuous circular aperture

The above-presented UCA can be considered as a spatially sampled (unbaffled) circular aperture using M sensors.

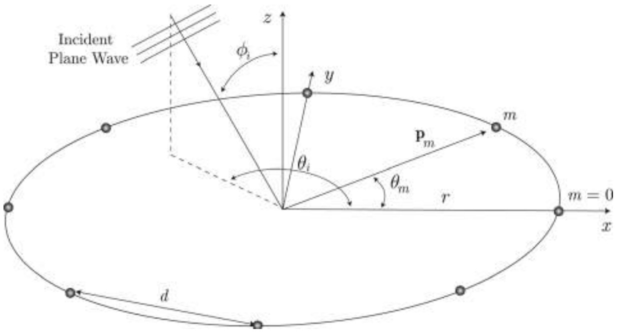

FIG. 1. Geometry of the UCA with M elements.

Assuming a plane wave impinging from (θ_i, $\phi_i = \pi/2$), the sound pressure at any point of a continuous circular aperture can be written in polar coordinates as

$$P(kr,\theta) = P_0 e^{jkr\cos(\theta-\theta_i)}, \tag{5}$$

where P_0 is the amplitude of the impinging wave. Note that the temporal term $e^{-j\omega t}$ has been suppressed for simplicity. Expanding the above-presented expression in a series of circular waves and after some mathematical treatment,[17] the incident pressure can be expressed as

$$P(kr,\theta) = P_0 \sum_{p=-\infty}^{\infty} j^p J_p(kr) e^{jp(\theta-\theta_i)}, \tag{6}$$

where J_p is a Bessel function of the first kind of order p. Note that the pressure in the aperture can be considered as a Fourier series and, therefore, it can be represented by

$$P(kr,\theta) = \sum_{p=-\infty}^{\infty} C_p e^{jp\theta}, \tag{7}$$

with Fourier coefficients (or circular harmonics) given by

$$C_p(kr,\theta_i) = P_0 j^p J_p(kr) e^{-jp\theta_i}. \tag{8}$$

In practice, continuous apertures must be sampled by means of a finite number of sensors. Section II B 2 describes the consequences of this sampling procedure.

2. Sampled circular aperture

The discretization of the continuous circular aperture by means of an UCA with M omnidirectional microphones results in the following Fourier coefficients:

$$\tilde{C}_p(kr) = \frac{1}{M}\sum_{m=0}^{M-1} \tilde{P}_m(kr) e^{-jp\theta_m}, \tag{9}$$

where $\tilde{P}_m$ is the measured sound pressure at the mth microphone (placed at angle θ_m). This sampling procedure implies an error in the Fourier coefficients as follows[11]:

$$\tilde{C}_p(kr,\theta_i) = C_p(kr,\theta_i) + \tilde{e}_p(kr,\theta_i), \tag{10}$$

$$\tilde{e}_p(kr,\theta_i) = P_0 \sum_{q=1}^{\infty}\left(j^q J_g(kr) e^{jg\theta_i} + j^h J_h(kr) e^{-jh\theta_i}\right), \tag{11}$$

where $g = Mq - p$ and $h = Mq + p$. Note that, according to Eq. (7), an infinite number of Fourier terms are needed to represent the sound pressure. In practice, the impinging wavefield must be decomposed into a maximum order L of circular harmonics and, thus, $M \geq 2L+1$. As a rule of thumb, $L \approx kr$ is usually chosen, since the value of a particular Bessel function is small when the order $p > 0$ exceeds the argument. The selection of an appropriate number of Fourier terms is further discussed in Sec. II D.

C. Beamforming

Modal beamforming aims at combining the different circular harmonic components (or phase modes) to form a beam with appropriate spatial selective properties. Ideally, the beamformer response should have a maximum when the beamformer is steered toward the source direction θ_i and should be zero in all other directions. This ideal response can be represented as a delta function as follows:

$$G_{\text{ideal}}(kr,\theta) = P_0\delta(\theta-\theta_i). \tag{12}$$

It can be shown that this ideal response is achieved by adding an infinite number of modes, so that the ideal beamformer can be written as[13]

$$G_{\text{ideal}}(kr,\theta) = \sum_{p=-\infty}^{\infty} \frac{C_p(kr,\theta_i)}{j^p J_p(kr)} e^{jp\theta}. \tag{13}$$

As discussed in the Sec. II B 2, when using a real UCA the number of modes must be truncated to a maximum order L. Moreover, the modal coefficients correspond to those of a sampled circular aperture, resulting in the following response:

$$G_{\text{CHB}}(kr,\theta) = \sum_{p=-L}^{L} \frac{\tilde{C}_p(kr,\theta_i)}{j^p J_p(kr)} e^{jp\theta}. \tag{14}$$

The output of the beamfomer for a steering direction θ_s can be expressed as

$$Y(kr,\theta_s) = \frac{1}{2L+1}\sum_{p=-L}^{L} \tilde{C}_p(kr)B_p(kr)H_p(\theta_s), \tag{15}$$

where B_p is an equalization factor given by

$$B_p(kr) = j^{-p}J_p^{-1}(kr) \tag{16}$$

and H_p is a frequency-independent phase alignment factor,

$$H_p(\theta_s) = e^{jp\theta_s}. \tag{17}$$

The normalization term, $1/(2L+1)$, is equal to the number of circular harmonics in the sum in order to keep unchanged the amplitude of the impinging plane-wave.

D. Mode selection and regularization

As discussed in Sec. II C, the filters $B_p(kr)$ are aimed at equalizing the responses of the individual eigenbeams, which depend on the Bessel function $J_p(kr)$. Figure 2 shows the magnitude of the four lowest-order ($p=0,\ldots,3$) Bessel functions of the first kind for different values of the argument kr. Note that for a given value of kr, there are only some orders (modes) with non-negligible contribution. As already explained, the rule of thumb is usually to select the maximum order L as

$$L = \lceil kr \rceil, \tag{18}$$

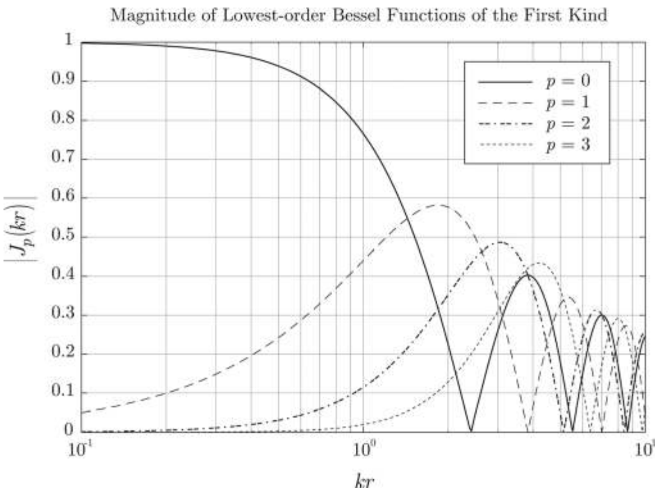

FIG. 2. Magnitude of the four lowest-order Bessel functions of the first kind.

where $\lceil\cdot\rceil$ is the ceiling function. Besides having orders with low magnitude, the different modes exhibit periodic zeros and, as a result, signals that carry components in the vicinity of the zeros cannot be completely resolved. To avoid this problem, the circular aperture can be mounted into a rigid cylindrical baffle[9,18]. However, it must be emphasized that, in this case, the array must be designed to have a height-to-radius ratio greater than 2.8 for approximating an ideal infinite-length cylinder.[18] Note that this physical requirement can be an issue in some practical applications (array 3 in Sec. IV would require a cylinder greater than 40.3 cm).

In order to avoid noise amplification due to large equalization values, Parthy *et al.*[9] proposed the use of Tikhonov-regularized filters, given by

$$B'_p(kr) = \frac{w_p^*(kr)}{\|w_p(kr)\|^2+\beta}, \tag{19}$$

where $w_p(kr) = B_p^{-1}(kr)$ and β is the regularization coefficient. The use of regularization, besides improving white noise gain, produces a smoother beampattern and provides increased robustness. In fact, directivity and robustness are linked to the value of β such that increasing β improves robustness and decreases directivity.

Figure 3 shows a comparison between the broadband beampatterns provided by conventional DSB (Ref. 19) and (regularized) CHB using a microphone array with $M=13$ and $r=0.12\,\text{m}$ steered to azimuth direction $\theta_s=\pi$. The selected regularization factor is $\beta = 6.5\times10^{-4}$. Note that CHB provides a narrower beampattern, although the effect of Bessel zeroes can be clearly seen in the response as vertical distortion lines around 1100, 1750, 2350 Hz, etc.

III. PROPOSED APPROACH

In the following we present our proposed approach where CHB is applied over a *T–F* processing framework to estimate the DOA of several source signals impinging simultaneously over an UCA.

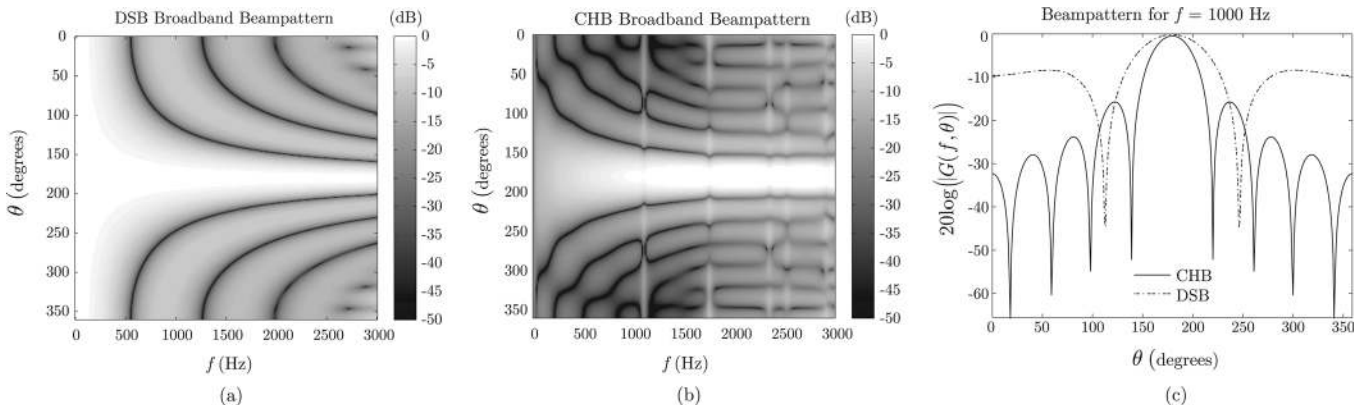

FIG. 3. Broadband beampatterns for steering direction $\theta_s = \pi$ using an UCA with $M = 13$ and $r = 0.12$ m. (a) Conventional delay and sum beamforming. (b) CHB beamforming. (c) Transversal section for $f = 1000$ Hz.

A. Signal model

Consider the following signal model with N sources s_n located in the far field at directions θ_n. For the sake of simplicity, an anechoic environment is assumed,

$$x_m(t) = \sum_{n=1}^{N} s_n(t - \delta_{mn}), \quad m = 0, \ldots, M-1, \tag{20}$$

where δ_{mn} is the time delay corresponding to the acoustic path between source n and microphone m.

In the short-time Fourier transform (STFT) domain, the above-presented signal model is expressed as

$$X_m(v,l) = \sum_{n=1}^{N} S_n(v,l) e^{-j2\pi f_v \delta_{mn}}, \tag{21}$$

where $X_m(v, l)$ and $S_n(v, l)$ are the *T–F* representations of the microphone signals and the sources, (v, l) are the frequency-bin and time-frame indices, and f_v is the analog frequency corresponding to frequency index v.

1. Speech sparsity in the T–F domain

Speech and music signals have been shown to be sparse in the *T–F* domain.[20] The probability density function of a sparse source has a peaky shape. This is due to the fact that the signal is close to zero at most *T–F* points and has large values in rare occasions. This property has been widely applied in many works related to source signal localization[21,22] and separation.[23,24] However, source sparsity alone is useless if the sources overlap to a high degree. The *disjointness* of a mixture of sources can be defined as the degree of non-overlapping of the mixed signals. An objective measure of disjointness is the so-called *W-Disjoint Orthogonality* (WDO).[25,26]

Spectral overlapping depends not only on source sparsity, but also on the mutual relationships between signals. Speech signals most often mix in a random and uncorrelated manner, such as in the cocktail party paradigm. Moreover, the disjointness properties of speech and music signals are dependent on the window size parameter, which affects the number of frequency bands in the analysis.[27] It is also worthwhile to remark that the sparsity and disjointness properties of audio signals become affected in reverberant environments. The room impulse response smears the energy in both time and frequency and so the spectral overlap between different sources in the *T–F* domain is increased with reverberation. Despite this effect, the assumption of non-overlapping sources has been shown to be still useful for sparsity-based applications such as source separation.[14]

Assuming that there is only one dominant source at each *T–F* bin (WDO assumption), the signal model can be further simplified as follows:

$$X_m(v,l) = S_{\breve{n}(v,l)}(v,l) e^{-j2\pi f_v \delta_{m\breve{n}(v,l)}}, \tag{22}$$

where $\breve{n}$ denotes the index of the dominant source at *T–F* point (v, l).

B. Time–frequency CHB

To estimate the direction of the dominant source signal at each *T–F* point, we perform CHB over each *T–F* element as follows. First, the microphone signals are transformed into the phase-mode domain at each *T–F* point (v, l) by taking the discrete Fourier transform (DFT) of $\mathbf{x}(v, l) = [X_0(v, l), \ldots, X_{M-1}(v, l)]^T$,

$$\mathbf{c}(v,l) = \mathrm{DFT}\{\mathbf{x}(v,l)\}. \tag{23}$$

The vector of coefficients must be accommodated to the following structure:

$$\mathbf{c}(v,l) = [\tilde{C}_0(f_v), \ldots, \tilde{C}_L(f_v), \tilde{C}_{-L}(f_v), \ldots, \tilde{C}_{-1}(f_v)]^T, \tag{24}$$

where the coefficients $\tilde{C}_p(f_v)$ correspond to those of a sampled circular aperture in Eq. (9). Note that, according to the rule $M \geq 2L+1$, the $(L+1)$th DFT coefficient must be discarded when having an even number of sensors.

Next, we define the following steering matrix, which is formed by Q different weighting vectors covering the azimuth range $\theta_q \in [0, 2\pi]$, $q = 1 \ldots Q$,

$$\mathbf{W} = [\mathbf{h}(\theta_1), \ldots, \mathbf{h}(\theta_Q)], \tag{25}$$

where

$$\mathbf{h}(\theta_q) = [e^{j0\theta_q}, ..., e^{jL\theta_q}, e^{j(-L)\theta_q}, ..., e^{j(-1)\theta_q}]^T. \quad (26)$$

Matrix $\mathbf{W}$ defines the angular range that will be spatially scanned for localizing the active sources. The equalization coefficient vector for each frequency is defined as

$$\mathbf{b}(v) = [B'_0(f_v), ..., B'_L(f_v), B'_{-L}(f_v), ..., B'_{-1}(f_v)]^T, \quad (27)$$

where the elements $B'_p(f_v)$ are computed as in Eq. (19) by using the relation $kr = (2\pi f_v/\mathrm{c})r$.

The equalized Fourier coefficients are calculated as

$$\bar{\mathbf{c}}(v, L) = \mathbf{c}(v, l) \circ \mathbf{b}(v), \quad (28)$$

where $\circ$ stands for the Hadamard product operator.

Finally, the beamformer output for each scanned angle $\mathbf{y}(v, l) = [Y(v, l, \theta_1), ..., Y(v, l, \theta_Q)]^T$ is computed by

$$y(v, l) = \frac{1}{2L+1}\mathbf{W}^T\bar{\mathbf{c}}(v, l). \quad (29)$$

C. DOA estimation

Assuming that there is one dominant sound source at each T–F point, the beamformer output will have maximum power at its corresponding arrival direction. Therefore, the DOA angle at each T–F bin can be estimated as

$$\hat{\theta}(v, l) = \arg\max_{\theta_q}\{|Y(v, l, \theta_q)|^2\}, \quad (30)$$

where

$$\hat{\theta}(v, l) \approx \theta_{\breve{n}(v,l)}, \quad (31)$$

$\theta_{\breve{n}(v,l)}$ being the real DOA corresponding to the dominant source at (v, l).

Since multiple simultaneous sources are dominant at different T–F points, the histogram of DOA estimates computed over the T–F plane shows clear peaks corresponding to the locations of the different sources. Although in this section an anechoic signal model has been assumed, in Sec. IV it will be shown how the method performs very robustly under adverse acoustic conditions including high reverberation and low SNR.

IV. EXPERIMENTS

The following is aimed at studying the performance of the proposed method (denoted in the following as TF-CHB) by considering different array configurations in diverse acoustic environments. First, acoustic simulations based on the image source method[28] are employed to analyze the influence of room reflections, noise, number of sources, and number of microphones. Then, the performance of the method is compared to other baseline techniques, namely conventional CHB, DSB, and super-resolution eigenbeamforming ESPRIT (EB-ESPRIT).[11,29] Localization performance is also analyzed with real recordings obtained from publicly available data.

A. Simulated recordings

The use of simulated recordings allows for understanding better how noise and reverberation affect localization accuracy for a given array configuration. In addition, synthetic recordings make it easier to observe which are the performance improvements obtained when using a higher number of microphones in a multi-source environment. Male and female speech fragments (4 s long) provided with the "*Dev2*" dataset of the *Signal Separation Evaluation Campaign*[30] were used as source signals. The sampling frequency used was $f_s = 8$ kHz, thus, all the simulated arrays were designed to have an aliasing frequency $f_{\mathrm{al}} = 4$ kHz. Three array configurations were considered:

(1) Array 1—$M = 5$, $r = 3.6$ cm.
(2) Array 2—$M = 9$, $r = 6.3$ cm.
(3) Array 3—$M = 21$, $r = 14.4$ cm.

Regarding the source arrangement, two complex multi-source cases were considered:

(1) Case 1—$N = 4$,

$\theta_n \in \{20°, 110°, 200°, 290°\}$.

(2) Case 2—$N = 8$,

$\theta_n \in \{20°, 65°, 110°, 155°, 200°, 245°, 290°, 335°\}$.

To evaluate the influence of room reflections, a box-shaped room with dimensions 6 m × 8 m × 3 m was simulated. The sources were located 2 m apart from the array center as shown in Fig. 4. The influence of reverberation was controlled by means of the wall reflection factor ρ,[31] which is assumed to be the same for all the room walls. Three different reflection factors were tested: $\rho = 0$ (anechoic), $\rho = 0.5$ (moderate reverberation), and $\rho = 0.9$ (high reverberation). Moreover, additive white Gaussian noise with power σ_m^2 is added to each microphone signal in order to provide different SNR values,

$$\mathrm{SNR} = 10\log\left(\frac{\left(\frac{1}{T}\right)\sum_t^T\sum_{n=1}^N s_{nm}^2(t)}{\sigma_m^2}\right), \quad (32)$$

where $s_{nm}(t)$ are the T sample's long original source signals convolved with the simulated source-to-microphone impulse responses, i.e., $s_{nm}(t) = s_n(t)*h_{mn}(t)$. The SNR values considered are SNR $= \infty$ dB (noise-free), SNR $= 10$ dB (noisy), and SNR $= 0$ dB (very noisy).

Regarding the processing parameters, STFTs were computed using Hamming windows of 512 samples length and 50% overlap. This value has been shown to be appropriate for speech signals.[20] The regularization factor was set to $\beta = 6.5 \times 10^{-4}$. The azimuthal space was scanned at 360 uniformly spaced angles ($Q = 360$), providing an angular resolution of 1°.

The resulting DOA histograms for $N = 4$ sources and $N = 8$ sources are shown in Figs. 5 and 6, respectively. Black dots in the θ axis denote the actual source locations. To ease

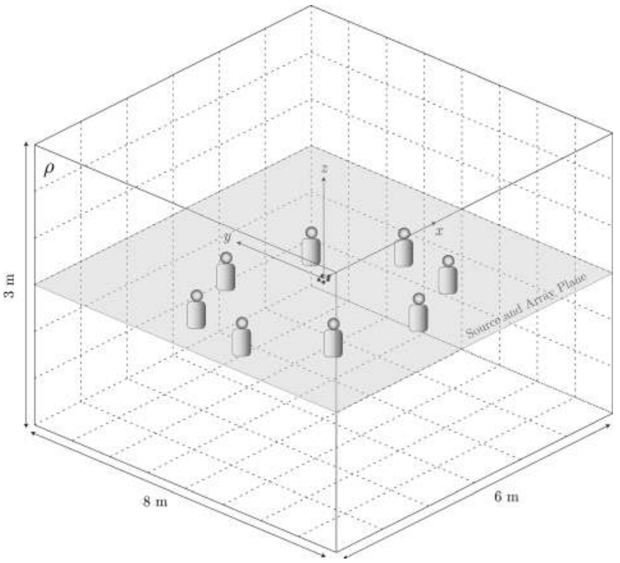

FIG. 4. Simulated room with dimensions 6 m × 8 m × 3 m and wall reflection factor ρ. The sources and the array are located on the xy plane.

the peak detection procedure, a low-pass filter that removes spurious peaks has been applied to the histograms, making them smoother. The final DOA of the sources can be obtained by using several alternatives, such as peak picking techniques or other model-based methods. In this paper, a simple peak picking method (findpeaks function in MATLAB) has been used to detect the local maxima in the smoothed histograms. The DOAs are assumed to be given by the N strongest peaks having a minimum separation distance of 10°. The root mean squared error (RMSE) values obtained for each array configuration are also presented in Table I. In the anechoic noise-free case [Figs. 5(a) and 6(a)], the histograms show very clear peaks located at the real source locations. Note that the resulting peak width is very dependent on the number of microphones, being considerably narrower if more elements are used. When noise and reverberation are present in the input signals, the histograms become noisier and the average peak width is substantially increased [Figs. 5(b)–5(i) and 6(b)–6(i)].

The effect of noise and reverberation in the performance can be critical depending on the number of simultaneous sources and their angular separation. Due to the increased peak width in adverse acoustic conditions, high-power sources located very close to low-power sources could mask the latter ones. Nevertheless, since adding more microphones results in narrower peak widths, localization accuracy can always be improved by using a higher number of microphones.

As shown in Table I, the localization accuracy for all the simulated cases depends on the above-described factors. It must be emphasized that a scenario with eight simultaneous speakers is a very extreme case in practice,

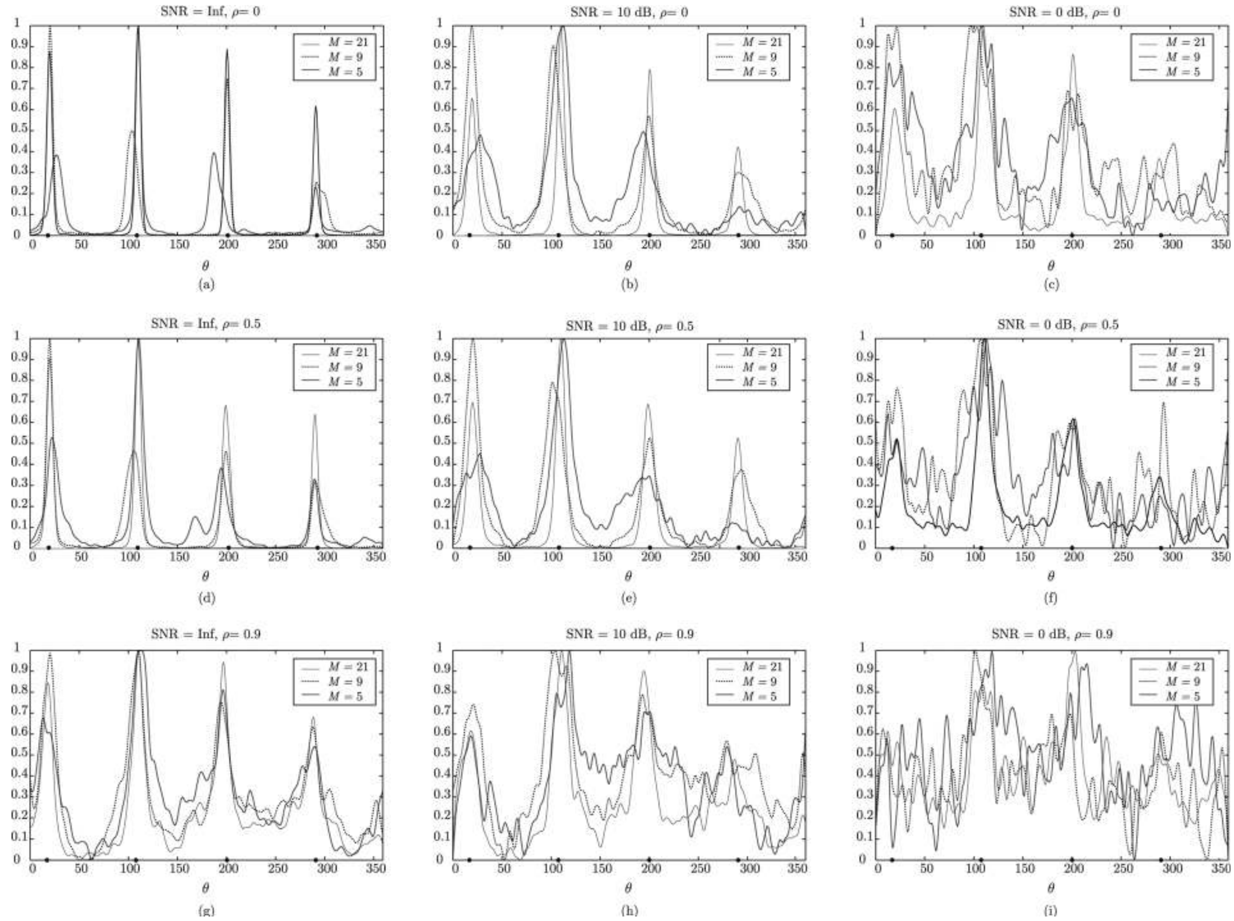

FIG. 5. Normalized histograms obtained for simulated recordings with $N = 4$ sources using different array configurations in diverse acoustic environments: (a) $\rho = 0$, SNR $= \infty$ dB, (b) $\rho = 0$, SNR $= 10$ dB, (c) $\rho = 0$, SNR $= 0$ dB, (d) $\rho = 0.5$, SNR $= \infty$ dB, (e) $\rho = 0.5$, SNR $= 10$ dB, (f) $\rho = 0.5$, SNR $= 0$ dB, (g) $\rho = 0.9$, SNR $= \infty$ dB, (h) $\rho = 0.9$, SNR $= 10$ dB, (i) $\rho = 0.9$, SNR $= 0$ dB.

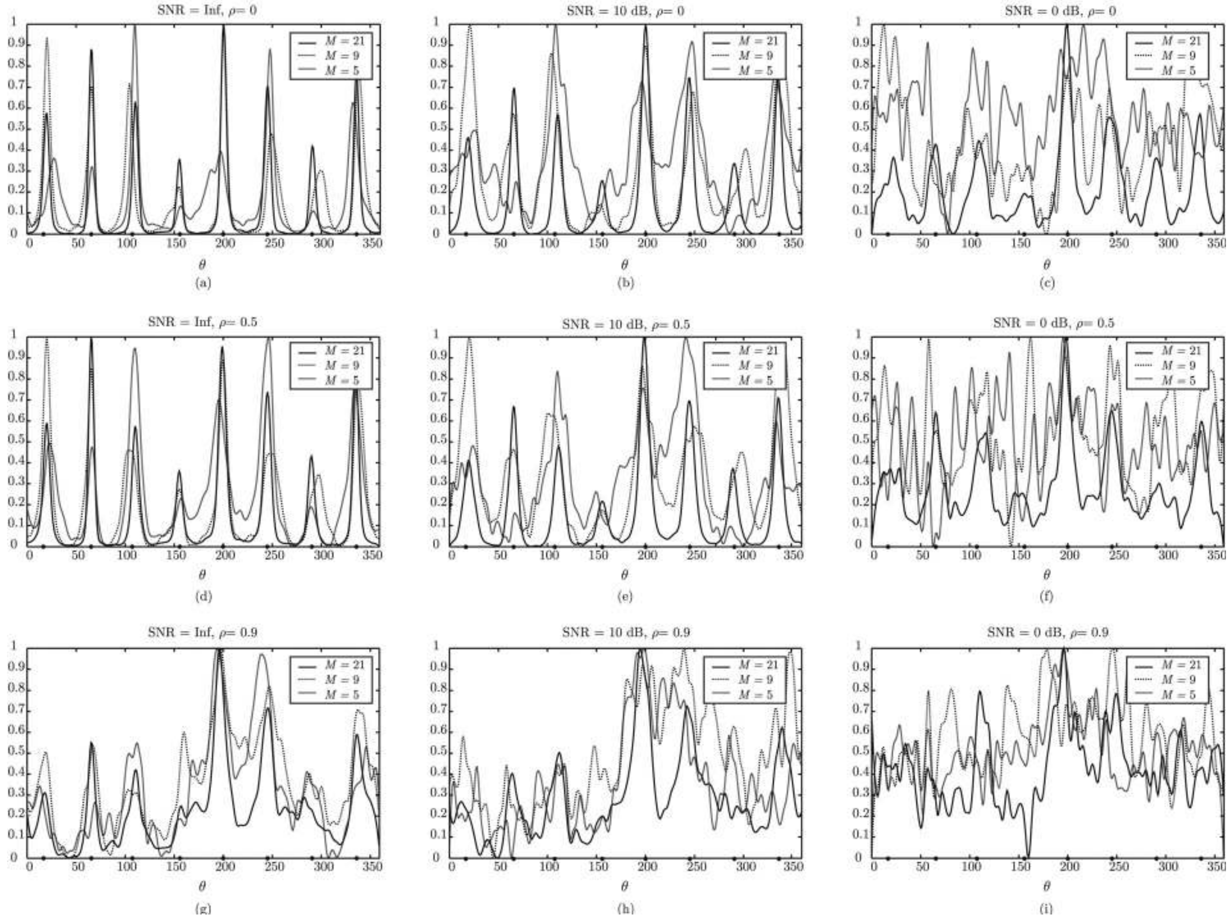

FIG. 6. Normalized histograms obtained for simulated recordings with $N=8$ sources using different array configurations in diverse acoustic environments: (a) $\rho=0$, SNR $=\infty$ dB, (b) $\rho=0$, SNR $=10$ dB, (c) $\rho=0$, SNR $=0$ dB, (d) $\rho=0.5$, SNR $=\infty$ dB, (e) $\rho=0.5$, SNR $=10$ dB, (f) $\rho=0.5$, SNR $=0$ dB, (g) $\rho=0.9$, SNR $=\infty$ dB, (h) $\rho=0.9$, SNR $=10$ dB, (i) $\rho=0.9$, SNR $=0$ dB.

since having more than two or three simultaneous sources in a speech communication environment (such as a teleconferencing or meeting room) is not usual.[32] In any case, note that the average localization error in moderate reverberation and noise conditions is around 1° for "array 3" and 8° for "array 2" when there are four simultaneous speakers.

1. Comparison with other methods

In the following, the performance of TF-CHB is compared to that of other well-known techniques. These techniques are the conventional DSB,[33] CHB,[13] and EB-ESPRIT.[11] All these methods are well-established localization techniques using circular arrays. DSB and CHB were compared by Tiana-Roig *et al.*,[13] where it was shown that CHB, despite being less robust in the presence of noise, has better angular resolution and sidelobe characteristics than DSB. The sound localization capabilities of modal arrays were also examined by Teutsch and Kellermann,[11] where the ESPRIT algorithm was applied over the phase-mode time-domain signals to localize several sources.

Figure 7 shows a comparison between the normalized angular output power of these methods and TF-CHB. Since EB-ESPRIT provides directly the directions of the estimated sources, the results for this method are represented as vertical lines at the estimated directions. The different panels show the results for case 1 and array 3 ($N=4$, $M=21$) in the same acoustic conditions as in Fig. 5. It can be clearly observed that the improved beam-width and sidelobe properties of CHB with respect to DSB results in narrower source peaks, thus, it provides better angular resolution. However, it

TABLE I. RMSE for different array configurations.

	$N=8$			$N=4$		
SNR (dB)	∞	10	0	∞	10	0
			Array 1			
$\rho=0.0$	6.60	9.35	14.31	5.39	9.22	11.76
$\rho=0.5$	11.29	13.86	25.86	8.60	10.25	12.38
$\rho=0.9$	19.45	27.65	33.54	11.84	14.76	19.72
			Array 2			
$\rho=0.0$	6.01	8.71	14.06	4.12	8.02	8.67
$\rho=0.5$	8.19	13.82	22.09	6.08	8.06	9.44
$\rho=0.9$	10.51	23.64	32.75	7.02	10.74	12.38
			Array 3			
$\rho=0.0$	0.00	1.41	6.04	0.00	1.00	3.35
$\rho=0.5$	1.00	2.45	8.25	1.00	1.00	3.35
$\rho=0.9$	8.61	14.56	26.08	3.74	6.48	9.91

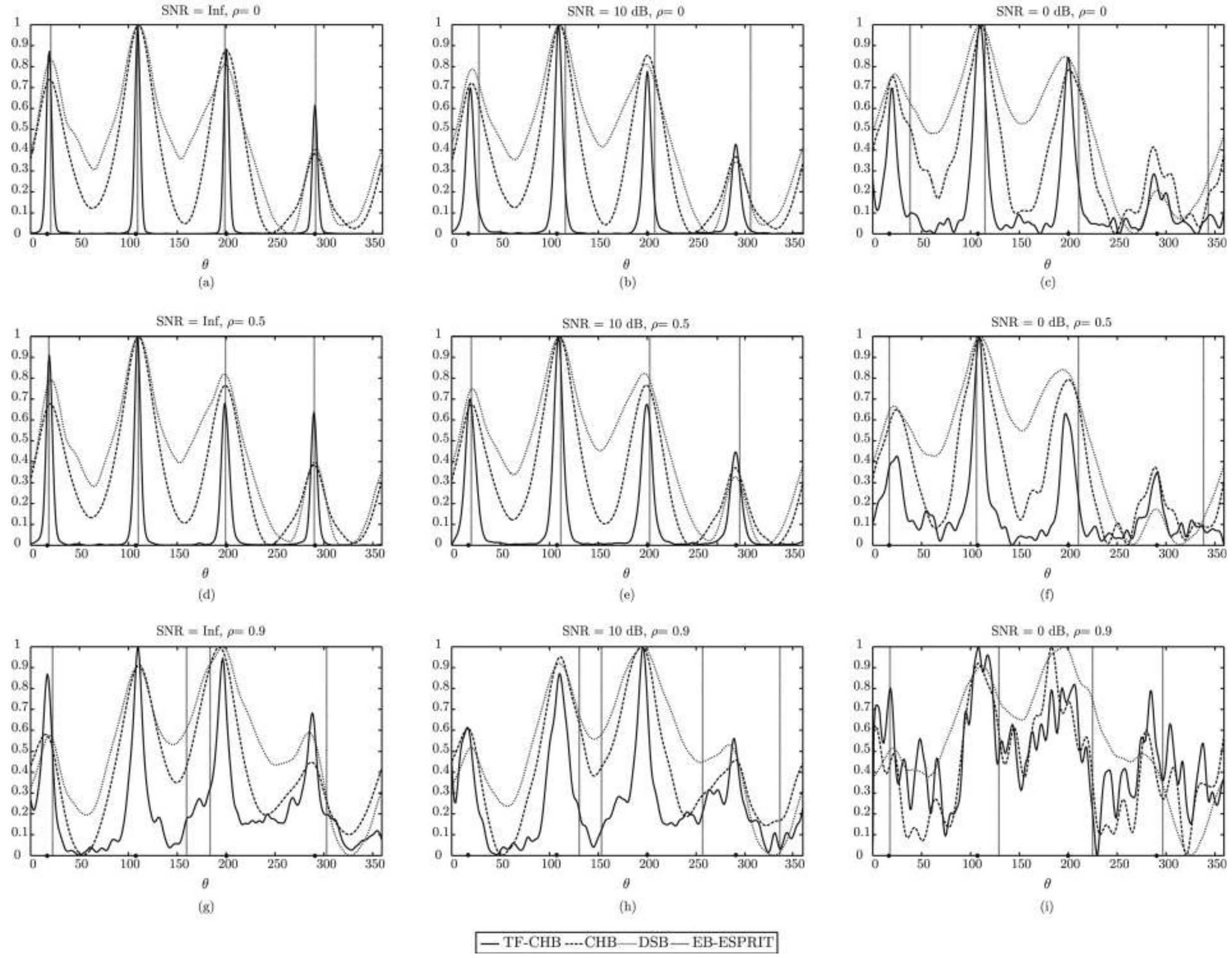

FIG. 7. Comparison between TF-CHB, DSB, CHB, and EB-ESPRIT: (a) $\rho = 0$, SNR $= \infty$ dB, (b) $\rho = 0$, SNR = 10 dB, (c) $\rho = 0$, SNR = 0 dB, (d) $\rho = 0.5$, SNR $= \infty$ dB, (e) $\rho = 0.5$, SNR = 10 dB, (f) $\rho = 0.5$, SNR = 0 dB, (g) $\rho = 0.9$, SNR $= \infty$ dB, (h) $\rho = 0.9$, SNR = 10 dB, (i) $\rho = 0.9$, SNR = 0 dB.

should also be noted that DSB is more robust than CHB in the presence of white noise, as shown in Fig. 7(i). EB-ESPRIT seems to be a very good method with moderate reverberation and noise levels, but its performance is severely degraded in extreme conditions. Table II shows the RMSE of these methods both for $N = 4$ and $N = 8$ speech sources. RMSE values for cases having non-observable peaks are not provided (the sources cannot be detected). Note that TF-CHB generally outperforms all the other methods, especially when there is a very high number of active sound sources. Moreover, it can be also observed how TF-CHB remains quite robust under very adverse acoustic conditions although, as expected, the localization accuracy decreases significantly.

B. Real recordings

Real data collected from the publicly available AV16.3 corpus[34] have been used to test our method with signals captured from a real UCA in a meeting room with three

TABLE II. RMSE for different localization methods using circular arrays.

	SNR = ∞				SNR = 10 dB				SNR = 0 dB			
	TF-CHB	CHB	DSB	EB-ESPRIT	TF-CHB	CHB	DSB	EB-ESPRIT	TF-CHB	CHB	DSB	EB-ESPRIT
						$N = 4$						
$\rho = 0.0$	0.00	0.00	2.06	1.42	1.00	0.00	2.24	2.54	3.35	2.06	4.12	6.95
$\rho = 0.5$	1.00	1.00	2.24	1.82	1.00	1.00	3.16	4.32	3.35	2.24	5.50	21.98
$\rho = 0.9$	3.74	5.20	4.03	43.30	6.48	8.79	5.83	43.39	9.91	14.08	14.76	53.52
						$N = 8$						
$\rho = 0.0$	0.00	5.01	4.24	23.80	1.41	6.09	5.91	31.13	6.04	13.21	12.89	65.27
$\rho = 0.5$	1.00	6.33	7.94	30.71	2.45	11.36	9.01	35.38	8.25	25.31	—	73.08
$\rho = 0.9$	8.61	—	—	50.40	14.56	—	—	52.82	26.08	—	—	73.59

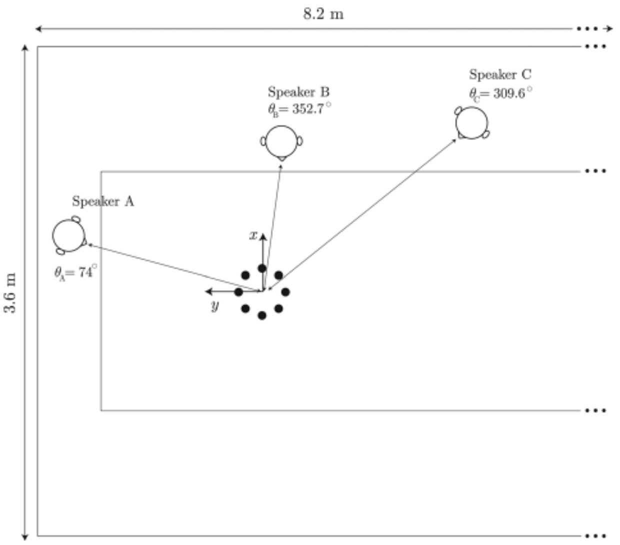

FIG. 8. Setup for the experiment with real recordings obtained from the AV16.3 public corpus.

simultaneous human speakers (see Fig. 8). This corpus has been widely used in many works related to speech processing.[35,36] Specifically, the signals used in this work correspond to the corpus recording labeled as "seq37-3 p-0001," using 9 of the 32 segmented speech fragments (three fragments for each case, $N=1$, $N=2$, and $N=3$). The recordings were collected in the IDIAP Smart Meeting Room,[37] with dimensions 3.6 m × 8.2 m × 2.4 m and an approximate reverberation time of $T_{60}=0.2$ s. An UCA with $M=8$ microphones and radius $r=10$ cm was used to capture the speech signals coming from three speakers located at different positions, as shown in Fig. 8. The sampling frequency of the original signals was $f_s=16$ kHz, however, the signals were resampled to 8 kHz in order to work with the same processing parameters as in Sec. IV A. The RMSE for each simultaneous talking case is presented in Table III. It is worthwhile to remark that the sources are real human speakers and, as opposed to loudspeaker sources, they tend to slightly change their head position as they speak. Therefore, due to these slight head movements, the obtained RMSE is not only a consequence of the localization method but also a side effect of the real application scenario.

V. DISCUSSION

The performance evaluation carried out in Sec. IV clearly shows how the localization accuracy achieved by the proposed method depends on the array design, the acoustic environment, and the source arrangement. Regarding array design considerations, two factors are important: The inter-microphone spacing and the number of microphones. While the first one limits the maximum working frequency, the second determines the robustness of the method under adverse acoustic conditions. Basically, accurate localization is possible for most practical situations if a sufficient number of microphones is used. The optimal number of microphones depends on the application scenario—the number of possible simultaneous sources, the accuracy needed, and the noise and/or reverberant characteristics of the room. For example, the eight-microphone array used in the experiment with real recordings has been shown to provide very good results in a common meeting environment.

TABLE III. RMSE for real recordings.

	$N=1$	$N=2$	$N=3$
	Speaker B	Speaker B, C	Speaker A, B, C
RMSE	1.18	2.0	2.98

It is also important to notice that the presented approach can be utilized together with other modeling techniques based on mixtures of distributions. As explained in Sec. IV, although in this paper a simple peak picking technique has been used to estimate the DOA of the sources, more sophisticated algorithms such as Gaussian Mixture Modelling[38] or Laplacian Mixture Modelling[7] can be applied to increase the robustness of the method when histogram peaks are not easily distinguishable.

VI. CONCLUSION

In this paper, a broadband acoustic source localization method based on *T–F* processing and modal beamforming has been proposed. A sparsity-motivated approach was presented to localize several simultaneous sound sources in adverse acoustic conditions. To this end, CHB with Tikhonov regularization is applied over each *T–F* point for steering the array toward a set of angles covering the azimuth plane. The angle with highest power is assumed to be a DOA estimate of the dominant source at each *T–F* point. Unlike other localization approaches working in the *T–F* domain, the proposed method exploits the use of the circular array geometry from a well-known modal processing framework. Meaningful experiments were conducted to evaluate the performance of the method under many different acoustic conditions and array configurations. Both simulated and real recordings were used. The results have shown that accurate localization performance can be achieved for most practical situations. In addition, the performance of the method has been compared to other baseline localization techniques based on UCA processing, showing the benefits of the proposed method. While noise and reverberation substantially affect localization performance, using a higher number of microphones allows one to increase the robustness of the method. Nevertheless, experiments in real situations have shown that a moderate number of microphones provides sufficient localization accuracy for most practical applications.

ACKNOWLEDGMENTS

The Spanish Ministry of Science and Innovation supported this work under Project No. TEC2009-14414-C03-01.

[1] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," J. Acoust. Soc. Am. **116**, 3075–3089 (2004).

[2] C. Liu, B. C. Wheeler, R. C. Bilger, C. R. Lansing, and A. S. Feng, "Localization of multiple sound sources with two microphones," J. Acoust. Soc. Am. **108**, 1888–1905 (2000).

[3] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, edited by M. S. Brandstein and D. Ward (Springer, Berlin, 2001), Chap. 8, pp. 157–180.

[4] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," Rob. Auton. Syst. J. **55**, 216–228 (2007).

[5] N. Madhu and R. Martin, *Advances in Digital Speech Transmission* (Wiley, New York, 2008), pp. 135–166.

[6] S. Haohai, H. Teutsch, E. Mabande, and W. Kellermann, "Robust localization of multiple sources in reverberant environments using EB-ESPRIT with spherical microphone arrays," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic (2011), pp.117–120.

[7] M. Cobos, J. J. Lopez, and D. Martinez, "Two-microphone multi-speaker localization based on a Laplacian mixture model," Digit. Signal Process. **21**, 66–76 (2011).

[8] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with arbitrarily arranged multiple sensors," J. Signal Process. Syst. **63**, 265–275 (2009). Available at http://www.springer.com/engineering/signals/journal/11265.

[9] A. Parthy, N. Epain, A. van Schaik, and T. J. Craig, "Comparison of the measured and theoretical performance of a broadband circular microphone array," J. Acoust. Soc. Am. **130**, 3827–3837 (2011).

[10] B. Rafaely, "Plane-wave decomposition of the sound field on a sphere by spherical convolution," J. Acoust. Soc. Am. **116**, 2149–2157 (2004).

[11] H. Teutsch and W. Kellermann, "Acoustic source detection and localization based on wavefield decomposition using circular microphone arrays," J. Acoust. Soc. Am. **120**, 2724–2736 (2006).

[12] H. Teutsch, *Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decomposition* (Springer, Berlin, 2007), pp. 150–188.

[13] E. Tiana-Roig, F. Jacobsen, and E. Fernandez Grande, "Beamforming with a circular microphone array for localization of environmental noise sources," J. Acoust. Soc. Am. **128**, 3535–3542 (2010).

[14] M. Cobos and J. J. Lopez, "Two-microphone separation of multiple speakers based on interclass variance maximization," J. Acoust. Soc. Am. **127**, 1661–1673 (2010).

[15] W. Zhang and B. D. Rao, "A two microphone-based approach for source localization of multiple speech sources," IEEE Trans. Audio, Speech, Lang. Process. **18**, 1913–1928 (2010).

[16] S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," J. Acoust. Soc. Am. **123**, 2136–2147 (2008).

[17] D. E. N. Davies, "Circular arrays," in *The Handbook of Antenna Design*, edited by A. W. Rudge, K. Milne, A. D. Olver, and P. Knight (Peregrinus, London, 1983), Vol. 2, Chap. 12, pp. 298–310.

[18] H. Teutsch, "Wavefield decomposition using microphone arrays and its application to acoustic scene analysis," Ph.D. thesis, Friedrich-Alexander-Universität, Erlangen-Nürnberg, Germany, 2005.

[19] G. Elko and J. Meyer, "Microphone arrays," in *Springer Handbook of Speech Processing*, edited by J. Benesty, M. Sondhi, and Y. Huang (Springer, Berlin, 2008), Chap. 50, pp. 1021–1042.

[20] J. J. Burred and T. Sikora, "On the use of auditory representations for sparsity-based sound source separation," in *Proceedings of the 5th International Conference on Information, Communications and Signal Processing (ICICS 2005)*, Bangkok, Thailand (2005).

[21] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Performance evaluation of sparse source separation and DOA estimation with observation vector clustering in reverberant environments," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Paris (2006).

[22] S. Rickard and F. Dietrich, "DOA estimation of many w-disjoint orthogonal sources from two mixtures using DUET," in *Proceedings of the 10th IEEE Workshop on Statistical Signal and Array Processing (SSAP2000)*, Pocono Manor, PA (2000), pp. 311–314.

[23] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. Signal Process. **52**, 1830–1847 (2004).

[24] P. Bofill and M. Zibulevski, "Underdetermined blind source separation using sparse representations," Signal Process. **81**, 2353–2362 (2001).

[25] S. Rickard and O. Yilmaz, "On the w-disjoint orthogonality of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, FL (2002), pp. 529–532.

[26] A. Jourjine, S. Richard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'00)*, Istanbul, Turkey (2000), Vol. 5, pp. 2985–2988.

[27] J. J. Burred, "From sparse models to timbre learning: New methods for musical source separation," Ph.D. thesis, Technical University of Berlin, 2008.

[28] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am. **65**, 943–950 (1979).

[29] H. Teutsch and W. Kellermann, "EB-ESPRIT: 2D localization of multiple wideband acoustic sources using eigen-beams," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2005)*, Philadelphia (2005), Vol. 3, pp. 89–92.

[30] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," Lect. Notes Comput. Sci. **5441**, 734–741 (2009).

[31] H. Kuttruff, *Room Acoustics* (Taylor & Francis, Abingdon, UK, 2000).

[32] J. Skowronek, A. Raake, K. Hoeldtke, and M. Geier, "Speech recordings for systematic assessment of multi-party conferencing," in *Proceedings of Forum Acusticum 2011* (European Acoustics Association, Aalborg, Denmark, 2011), pp. 111–116.

[33] H. Krim and M. Viberg, "Two decades of array signal processing research, the parametric approach," IEEE Signal Process. Mag. **13**, 67–94 (1996).

[34] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16.3: An audio-visual corpus for speaker localization and tracking," in *Proceedings of the 1st International Workshop on Machine Learning for Multimodal Interaction (MLMI 2004)*, Martigny, Switzerland (2004), pp. 192–195.

[35] G. Lathoud and M. Magimai-Doss, "A sector-based, frequency-domain approach to detection and localization of multiple speakers," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia (2005), Vol. 3, pp. 265–268.

[36] A. Plinge, M. H. Hennecke, and G. A. Fink, "Robust neuro-fuzzy speaker localization using a circular microphone array," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel Aviv, Israel (2010).

[37] D. C. Moore, "The IDIAP smart meeting room," Technical Report IDIAP-COM 02-07, IDIAP (2002), available at http://glat.info/ma/av16.3/com02-07.pdf (Last viewed 4/5/2012).

[38] N. Madhu and R. Martin, "A scalable framework for multiple speaker localization and tracking," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle (2008).