MAY 04 2012

A directionally tunable but frequency-invariant beamformer on an acoustic velocity-sensor triad to enhance speech perception [a)]

Yue Ivan Wu; Kainam Thomas Wong; Xin Yuan; Siu-kit Lau; Shiu-keung Tang

J. Acoust. Soc. Am. 131, 3891–3902 (2012)
https://doi.org/10.1121/1.3701991

A directionally tunable but frequency-invariant beamformer on an acoustic velocity-sensor triad to enhance speech perception[a)]

Yue Ivan Wu
Centre for Signal Processing, Nanyang Technological University, 50 Nanyang Avenue, 639798 Singapore, Singapore

Kainam Thomas Wong[b)] and Xin Yuan
Department of Electronic and Information Engineering, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

Siu-kit Lau
Charles W. Durham School of Architectural Engineering and Construction, University of Nebraska–Lincoln, 1110 S. 67th Street, Omaha, Nebraska 68182-0816

Shiu-keung Tang
Department of Building Services Engineering, Hong Kong Polytechnic University, CD 637 Hung Hom, Kowloon, Hong Kong

(Received 9 August 2011; revised 20 March 2012; accepted 23 March 2012)

Herein investigated are computationally simple microphone-array beamformers that are independent of the frequency-spectra of all signals, all interference, and all noises. These beamformers allow the listener to tune the desired azimuth-elevation "look direction." No prior information is needed of the interference. These beamformers deploy a physically compact triad of three collocated but orthogonally oriented velocity sensors. These proposed schemes' efficacy is verified by a jury test, using simulated data constructed with Mandarin Chinese (a.k.a. Putonghua) speech samples. For example, a desired speech signal, originally at a very adverse signal-to-interference-and-noise power ratio (SINR) of −30 dB, may be processed to become fully intelligible to the jury. © *2012 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.3701991]

PACS number(s): 43.60.Fg, 43.60.Dh, 43.60.Gk [SAF] Pages: 3891–3902

I. INTRODUCTION

A. Microphone array receiver

Acoustic receivers need to function in adverse environments despite hardware limitations in the microphone transducer and in the signal-processing electronics. Microphone-reception technology would ideally isolate the desired sound signal (often speech) and would ideally suppress all undesired background noises (including other speakers, music, and other household noises). Complicating the situation is that these undesired noises are generally *a priori* unknown, uncontrollable, and unpredictable and that such interference typically overlaps with the desired speech signal, in *time* and *frequency*. Nevertheless, the *spatial* dimension could be exploited if the receiver deploys *multiple* microphones instead of a single microphone. By deploying an array of microphones (instead of one microphone), a signal-processing algorithm can electronically form spatial beams to pass the desired speaker, but spatial nulls other directions at which the dominant interferences impinge.[1]

However, the above-mentioned beamformer schemes are computationally complex in real time and require expensive and bulky electronic hardware. This real-time computational complexity arises from the following factors: (a) The beamformer weights vary with frequency due to the intersensor spatial phase factor across spatially displaced sensors. (b) Speech signals and most background noises are broadband, spanning over wide spectra of frequencies that are typically *a priori* unknown and time-varying. Due to (a) and (b), an array of displaced microphones needs to have its broadband acoustic data algorithmically decompose in real time, into a spectrally contiguous set of narrowband signals, each at a different frequency, then to be separately processed in real time by the hearing-aid electronics. Present microphone-array receivers thus require heavy real-time computations due to the microphone-array's intrinsic dependence on the incident source's frequency, bandwidth, and location in the near field versus the far field.

All above frequency-related complications can be avoided by using a different kind of acoustic sensor that will be presented herein—the acoustic velocity-sensor triad, which samples the incident acoustic wavefield not as a pressure *scalar* but as a particle-velocity *vector*.

B. The acoustic velocity-sensor triad—a.k.a. the "acoustic vector sensor" or the "vector hydrophone"

Customary microphones treat the incident acoustic wavefield as a scalar field, i.e., the acoustic "pressure" scalar, which

[a)]Part of this work was presented in the International Conference on Networked Sensing Systems, held in Penghu, Taiwan, on June 12–15, 2011 [K. T. Wong, Y. I. Wu, X. Yuan, S. k. Lau, and S. k. Tang, "A directionally tunable but frequency-invariant beamformer for an acoustic velocity-sensor triad to enhance speech perception," in *International Conference on Networked Sensing Systems* (June 12–15, 2011)].

[b)]Author to whom correspondence should be addressed. Electronic mail: ktwong@ieee.org

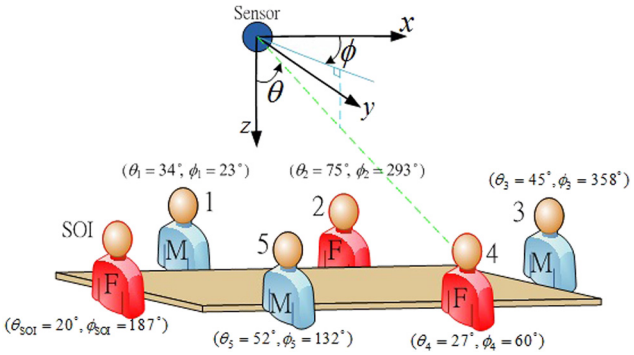

FIG. 1. (Color online) The simulated conference setting: Here, the "SOI" indicates the speaker of interest. The numbers 1–5 indices the interfering speakers. The "M" signifies a male speaker, whereas an "F" refers to a female speaker.

varies over time and space to form a scalar field. Thus overlooked is much information in the underlying acoustic "particle velocity vector"—a three-dimensional vector representing the pressure field's three partial derivatives with respect to the three Cartesian spatial coordinates. To measure any one Cartesian component of this vector, an acoustic particle-*velocity* sensor may be deployed along that Cartesian axis.

To treat the acoustic wavefield as a vector field (i.e., the particle-velocity field) and not merely as a scalar field (i.e., pressure field), all three Cartesian components of the particle-velocity vector are to be distinctly measured. That would allow beamforming over this acoustic particle-velocity vector to attain reception-diversity with respect to the azimuth-elevation direction of arrival (DOA), so as to enhance the signal of interest and to null the interfering signals. To facilitate this distinct processing of all three Cartesian components of the particle-velocity vector, the acoustic vector sensor (a.k.a. vector hydrophone) is available, which consists of three identical, but orthogonally oriented, acoustic *velocity* sensors (sometimes with an optional pressure sensor)[2]—all spatially co-located in a point-like geometry.

More mathematically, an acoustic *vector* sensor (placed at the origin of the three-dimensional Cartesian coordinates) would have this 3×1 array-manifold[3–5] in response to a unit-power incident acoustic wave that has traveled through an homogeneous isotropic medium from either a near-field or far-field emitter:

$$\mathbf{a}(\theta, \phi) \stackrel{\text{def}}{=} \begin{bmatrix} u(\theta, \phi) \\ v(\theta, \phi) \\ w(\theta) \end{bmatrix} \stackrel{\text{def}}{=} \begin{bmatrix} \sin(\theta)\cos(\phi) \\ \sin(\theta)\sin(\phi) \\ \cos(\theta) \end{bmatrix}, \quad (1)$$

where $\theta \in [0, \pi]$ signifies the elevation-angle measured from the positive z axis, $\phi \in [0, 2\pi)$ symbolizes the azimuth-angle measured from the positive x axis, $u(\theta, \phi)$ denotes the direction cosine along the x axis, $v(\theta, \phi)$ refers to the direction-cosine along the y axis, and $w(\theta)$ represents the direction-cosine along the z axis. The first, second, and third component of $\mathbf{a}(\theta, \phi)$ corresponds to the acoustic *velocity* sensor aligned along the x, y, and z axes, respectively. This collocated unit is intrinsically directional, potentially able to pick up only sounds arriving from a certain fixed azimuth-elevation direction, while suppressing noises and interfering sounds from other directions. The acoustic vector sensor's beam pattern and directivity have been investigated.[6–21]

This acoustic *vector*-sensor concept is practical. It has been implemented in hardware in various forms for air-acoustic applications. Acoustic vector sensors are commercially available. Acoustic vector sensors have undergone in-building room trials or atmospheric trials. The details are available from literature surveys[22–24] of the velocity sensor and the vector sensor, their hardware implementation, and their field trials.

It is essential to note that $\mathbf{a}(\theta, \phi)$ is independent of the frequency spectrum of the incident signal—independent of both the signal's frequency band and its time-frequency structure. Hence, *two*-dimensional azimuth-elevation spatial beamforming may be realized via $\mathbf{a}(\theta, \phi)$, *without* regard to the sources' frequency bands and frequency spectra and *without* regard to the sources' locations in the near field or the far field of the receiving acoustic vector sensor. It is precisely such frequency-*independence* that is lacking in any customary array of spatially displaced pressure microphones. That frequency dependency is exactly what renders a customary array of pressure microphones to have computationally complicated beamforming. Beamforming with an acoustic vector sensor has previously been investigated.[14,17,21,25–34]

C. Overview of the present investigation

This work verifies the beamforming efficacy of an acoustic vector sensor in a conferencing scenario whereby the receiver aims to isolate one speaker, while suppressing interfering speakers elsewhere in the room at unknown arbitrary locations. This spatial beamforming is to be performed with *no* prior knowledge of any time-frequency structure of any speaker. The desired DOA may be tuned by the user him/her/itself. The resulting beamformer outputs are clinically assessed by a jury against the corresponding speech

TABLE I. The speakers and their speech signals.

Source	DOA	Speaker's gender	Contents
SOI	$\{\theta_{SOI}, \phi_{SOI}\} = \{20^\circ, 187^\circ\}$	female	A book-reading in .mp3
Interfering speaker 1	$\{\theta_1, \phi_1\} = \{34^\circ, 23^\circ\}$	male	A book-reading in .mp3
Interfering speaker 2	$\{\theta_2, \phi_2\} = \{75^\circ, 293^\circ\}$	female	A news report in .rm
Interfering speaker 3	$\{\theta_3, \phi_3\} = \{45^\circ, 358^\circ\}$	male	A news report in .mp4
Interfering speaker 4	$\{\theta_4, \phi_4\} = \{27^\circ, 60^\circ\}$	female	A news report in .mpg
Interfering speaker 5	$\{\theta_5, \phi_5\} = \{52^\circ, 132^\circ\}$	male	A news report in .rm

samples before beamforming. All speakers and all jury members here are native speakers of Mandarin Chinese.

The rest of this paper is organized as follows: Sec. II A will describe the mathematical models for measurements collected by a single omni-directional microphone. Section II B will do the same for measurements from an acoustic vector sensor. Section II C will describe the multispeaker conferencing scenario in the Monte Carlo simulation. Section III will describe the "spatial matched filter" (SMF) beamformer, which can serve as a performance benchmark. Section IV will define the algorithmic steps in the proposed "minimum-power distortionless response" (MPDR) beamformer. Section V will discuss situations that have no perfect/prior tuning to the desired speaker. Section VI will describe the jury assessment. Section VII will conclude the entire paper.

II. THE MEASUREMENT MODEL

Denote the desired speaker's signal as $\sqrt{\mathcal{P}_{\mathrm{SOI}}}s_{\mathrm{SOI}}(t)$ with $\|s_{\mathrm{SOI}}(t)\|_2 = 1$, where $\|\cdot\|$ symbolizes the Frobenius norm over the entire observation duration. Symbolize the ith interfering speaker's signal as $\sqrt{\mathcal{P}_i}s_i(t)$, with $\|s_i(t)\|_2 = 1$.

A. Data measured by a isotropic sensor

An isotropic microphone would collect the following scalar datum at time t,

$$z_{\mathrm{ISO}}(t) = \sqrt{\mathcal{P}_{\mathrm{SOI}}}s_{\mathrm{SOI}}(t) + \sum_{i=1}^{I}\sqrt{\mathcal{P}_i}s_i(t) + \sqrt{\mathcal{P}_{\mathrm{n}}}n_{\mathrm{ISO}}(t), \tag{2}$$

with the real-value additive noise at time t being $\sqrt{\mathcal{P}_{\mathrm{n}}}n_{\mathrm{ISO}}(t)$ with $\|n_{\mathrm{ISO}}(t)\|_2 = 1$.

For subsequent discussion, define the signal-to-interference-plus-noise ratio (SINR) at the beamformer's *input* as

$$\mathrm{SINR}_{\mathrm{in}} = \frac{\mathcal{P}_{\mathrm{SOI}}}{\sum_{i=1}^{I}\mathcal{P}_i + \mathcal{P}_{\mathrm{n}}}. \tag{3}$$

B. Data measured by an acoustic vector sensor

An acoustic vector sensor, at time t, would collect the 3×1 data-vector,

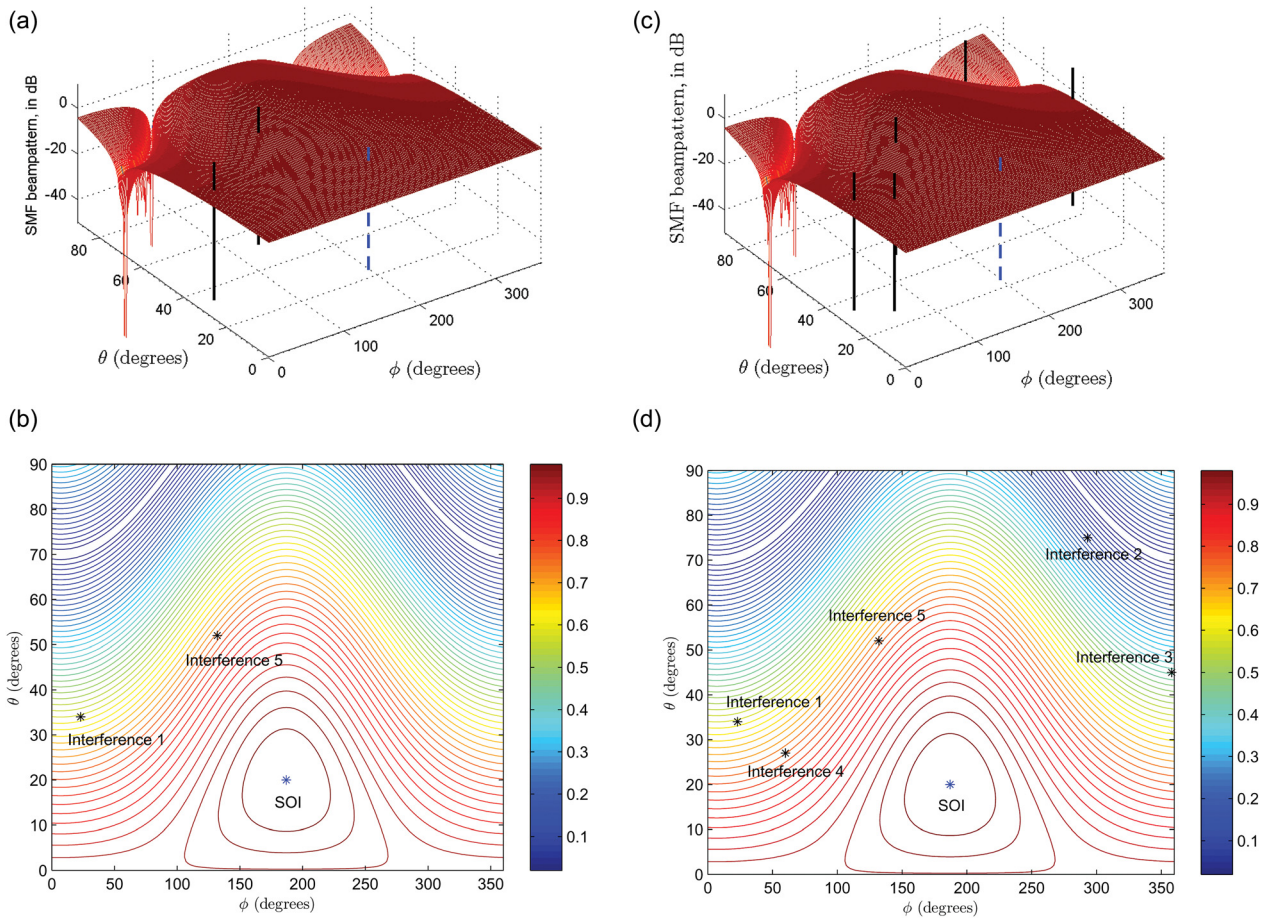

FIG. 2. (Color online) (a) The acoustic vector sensor's SMF beam pattern with three simultaneous speakers (the SOI, plus interfering speakers 1 and 5) in the conference-room setting of Fig. 1. Here SNR = 30 dB and INR = 33 dB. (b) The contour map of the SMF beam pattern with three simultaneous speakers (the SOI, plus interfering speakers 1 and 5) in the conference-room setting of Fig. 1. Here SNR = 30 dB and INR = 33 dB. (c) The acoustic vector sensor's SMF beam pattern simultaneously with all six speakers in the conference-room setting of Fig. 1. Here SNR = 30 dB and INR = 37 dB. (d) The contour map of the SMF beam pattern simultaneously with all six speakers in the conference-room setting of Fig. 1. Here SNR = 30 dB and INR = 37 dB.

$$\mathbf{z}_{\mathrm{AVS}}(t) = \sqrt{\mathcal{P}_{\mathrm{SOI}}}\mathbf{a}(\theta_{\mathrm{SOI}}, \phi_{\mathrm{SOI}})s_{\mathrm{SOI}}(t) + \underbrace{\sum_{i=1}^{I} \sqrt{\mathcal{P}_i}\mathbf{a}(\theta_i, \phi_i)s_i(t) + \sqrt{\mathcal{P}_{\mathrm{n}}}\mathbf{n}_{\mathrm{AVS}}(t)}_{=\mathbf{z}_{\mathrm{I+N}}(t)} \quad (4)$$

where $(\theta_{\mathrm{SOI}}, \phi_{\mathrm{SOI}})$ represents the elevation angle and the azimuth angle of the desired speaker relative to the microphone. Similarly, (θ_i, ϕ_i) denotes the corresponding angle of arrival of the ith interfering speaker. Please see Fig. 1. Moreover, the 3×1 real-value additive noise $\sqrt{\mathcal{P}_{\mathrm{n}}}\mathbf{n}_{\mathrm{AVS}}(t)$ needs *not* be spatiotemporally white; however, each of its entry has the same temporal statistics as $\mathbf{n}_{\mathrm{ISO}}(t)$ of Sec. II A.

The sample covariance matrix, based on data collected at $\{t = t_m, m = 1,\ldots, M\}$, may be expressed as

$$\mathbf{R}_{\mathrm{AVS}} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{z}_{\mathrm{AVS}}(t_m)\mathbf{z}_{\mathrm{AVS}}^T(t_m), \quad (5)$$

where the superscript T symbolizes the transposition operator.

For any arbitrary beamformer weight **w,** its enhancement of the SINR may be measured by its "array gain" (Ref. 35), defined as

$$\mathrm{G}(\mathbf{w}) = \frac{\mathrm{SINR}_{\mathrm{out}}(\mathbf{w})}{\mathrm{SINR}_{\mathrm{in}}} = \frac{|\mathbf{w}^T\mathbf{a}(\theta_{\mathrm{SOI}}, \phi_{\mathrm{SOI}})|^2}{\mathbf{w}^T\rho_{\mathrm{I+N}}\mathbf{w}}, \quad (6)$$

where

$$\mathrm{SINR}_{\mathrm{out}}(\mathbf{w}) = \frac{\mathcal{P}_{\mathrm{SOI}}|\mathbf{w}^T\mathbf{a}(\theta_{\mathrm{SOI}}, \phi_{\mathrm{SOI}})|^2}{\mathbf{w}^T\mathbf{R}_{\mathrm{I+N}}\mathbf{w}}, \quad (7)$$

denotes the output-SINR for the beamforming-weight vector **w,** with $\mathbf{R}_{\mathrm{I+N}} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{z}_{\mathrm{I+N}}(t_m)\mathbf{z}_{\mathrm{I+N}}^T(t_m)$ and $\rho_{\mathrm{I+N}} = \frac{\mathbf{R}_{\mathrm{I+N}}}{\sum_{i=1}^{I}\mathcal{P}_i + \mathcal{P}_{\mathrm{n}}}$.

C. The simulated "conference" setting

A conferencing setting will be simulated with speakers seated around a table, and a sensor mounted on the ceiling above the table. Figure 1 illustrates the three-dimensional spatial geometry among the speaker of interest (SOI), five interfering speakers, and the sensor. Without loss of generality, the sensor location constitutes the origin of the elevation-azimuth spherical coordinates, (θ, ϕ).

In subsequent simulations: Marked on Fig. 1 is each speaker's incident angle upon the sensor. Each speaker's DOA and

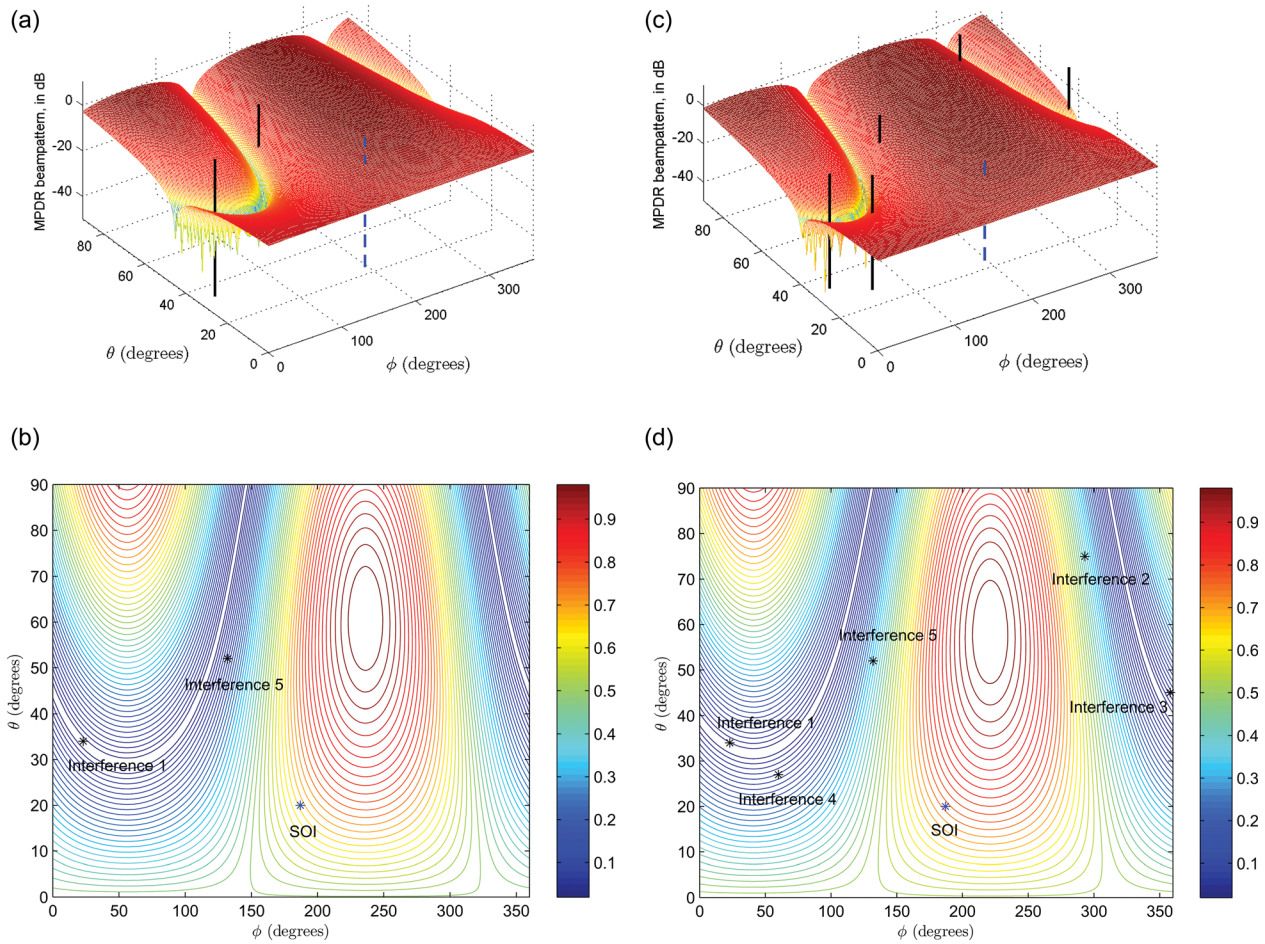

FIG. 3. (Color online) (a) The acoustic vector sensor's MPDR beam pattern with three simultaneous speakers (i.e., the SOI, plus interfering speakers 1 and 5) in the conference-room setting of Fig. 1. Here SNR = 30 dB and INR = 33 dB. (b) The contour map of the MPDR beam pattern with three simultaneous speakers (i.e., the SOI, plus interfering speakers 1 and 5) in the conference-room setting of Fig. 1. Here, SNR = 30 dB and INR = 33 dB. (c) The acoustic vector sensor's MPDR beam pattern simultaneously with all six speakers in the conference-room setting of Fig. 1. Here, SNR = 30 dB and INR = 37 dB. (d) The contour map of the MPDR beam pattern simultaneously with all six speakers in the conference-room setting of Fig. 1. Here SNR = 30 dB and INR = 37 dB.

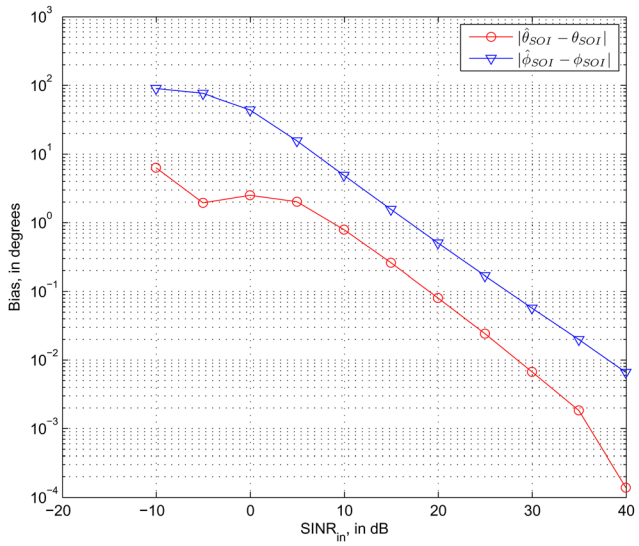

FIG. 4. (Color online) The acoustic vector sensor's self-tuning DOA-estimation bias with three simultaneous speakers. Here MUSIC is used and INR = 20 dB.

other details are listed in Table I. Each active speaker emits a Mandarin Chinese speech signal of bandwidth 44.1 kHz, downloaded from http://mp3.baidu.com/m?tn=baidump3, then converted into a "wave" file at the same bit-rate.

The individual speaker's speech samples are mixed and processed according to Eq. (2) as input to the single isotropic microphone and according to Eq. (4) as input to the acoustic vector-sensor beamformer. In all cases, all interfering signals are set to the same power, i.e., $\mathcal{P}_i$ are the same for all i. For subsequent analysis, define SNR $= \mathcal{P}_{SOI}/\mathcal{P}_n$ and INR $= \Sigma_{i=1}^{I} \mathcal{P}_i/\mathcal{P}_n = I\mathcal{P}_i/\mathcal{P}_n$; hence, $\text{SINR}_{in} = \text{SNR}/(\text{INR} + 1)$.

III. METHOD 1: SMF BEAMFORMER FOR THE TUNABLE ACOUSTIC VECTOR SENSOR

Suppose the user manually tunes the receiver toward the desired speaker. This would electronically produce a SMF (SMF) beamforming-weight vector $\mathbf{w}_{SMF}(\theta_{SOI}, \phi_{SOI}) = \mathbf{a}(\theta_{SOI}, \phi_{SOI})$ for the acoustic vector sensor, resulting in a beamformer output of

$$b_{tuned}(t) = [\mathbf{w}_{SMF}(\theta_{SOI}, \phi_{SOI})]^T \mathbf{z}_{AVS}(t) = \mathbf{a}^T(\theta_{SOI}, \phi_{SOI})\mathbf{z}_{AVS}(t). \quad (8)$$

The preceding beamforming-weight vector $\mathbf{w}_{SMF}(\theta_{SOI}, \phi_{SOI})$ is matched to the desired source's steering vector $\mathbf{a}(\theta_{SOI}, \phi_{SOI})$ but requires no prior knowledge of (θ_i, ϕ_i), $\forall_i$ of the interfering speakers/sources. This SMF beamformer is computationally very simple, requiring only three real-value multiplications per time-sampling instant.

Figures 2(a) and 2(b), respectively, show the SMF beamformer's azimuth-elevation beam pattern (i.e., $|[\mathbf{a}(\theta, \phi)]^T \mathbf{a}(\theta_{SOI}, \phi_{SOI})|$) and the corresponding contour map, for the conference-room setting in Fig. 1 with three simultaneous speakers, namely, the SOI, interfering speaker 1, and interfering speaker 5. The dashed line in Fig. 2(a) indicates the SOI's DOA upon the acoustic vector sensor, whereas the solid lines are the counterparts for the two interferences. Figures 2(c) and 2(d) are similar but with six simultaneous speakers' locations. All these figures clearly show that the SMF beam pattern does peak at the SOI, but the SMF beam pattern's null can mismatch most interfering sources.

For the three-speaker scenario in Fig. 1 (i.e., the SOI and interfering speakers 1 and 5) at INR = 10 dB, the SMF beamformer would attain an array gain of $G(\mathbf{w}_{SMF}(\theta_{SOI}, \phi_{SOI})) = 3.7$ dB.

IV. METHOD 2: MPDR BEAMFORMER FOR THE TUNABLE ACOUSTIC VECTOR SENSOR

The customary MVDR beamformer (a.k.a. the Capon beamformer)[36,37] minimizes the beamformer output power, while pre-serving the incident power (whether from the SOI and/or the interference and/or noises) from a desired DOA. The MVDR beamformer is linearly constrained to ensure no distortion at the specified "look direction" of $(\theta_{tune}, \phi_{tune})$ but to minimize the beamformer's overall output power. The MVDR-beamformer weight vector equals

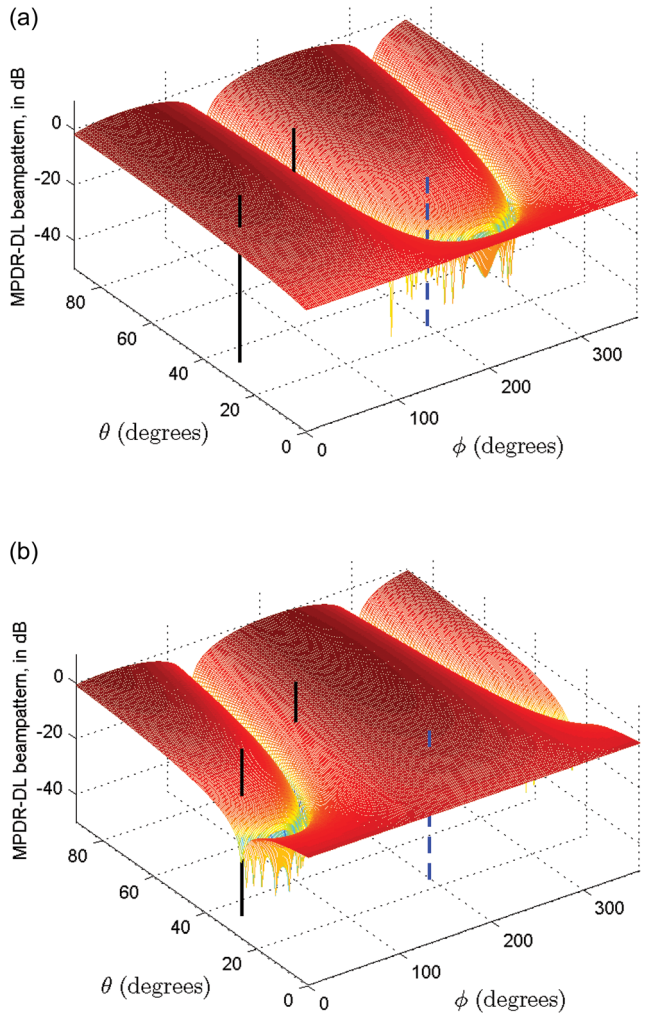

FIG. 5. (Color online) (a).The MPDR-DL beampattern with $\mathcal{P}_\ell = 0$, and with three speakers in Fig. 1 (i.e., the SOI and interfering speakers 1 and 5). Here, the pointing error equals $(\theta_{tune} - \theta_{SOI}, \phi_{tune} - \phi_{SOI}) = (15°, 15°)$, SNR = 30 dB and INR = 10 dB. (b) The MPDR-DL beampattern with $\mathcal{P}_\ell = 316.2$, and with three speakers in Fig. 1 (i.e., the SOI and interfering speakers 1 and 5). Here the pointing error equals $(\theta_{tune} - \theta_{SOI}, \phi_{tune} - \phi_{SOI}) = (15°, 15°)$, SNR = 30 dB and INR = 10 dB.

$$\mathbf{w}_{\mathrm{MVDR}}(\theta_{\mathrm{tune}},\phi_{\mathrm{tune}})=\frac{\mathbf{R}_{\mathrm{I+N}}^{-1}\mathbf{a}(\theta_{\mathrm{tune}},\phi_{\mathrm{tune}})}{\mathbf{a}^{T}(\theta_{\mathrm{tune}},\phi_{\mathrm{tune}})\mathbf{R}_{\mathrm{I+N}}^{-1}\mathbf{a}(\theta_{\mathrm{tune}},\phi_{\mathrm{tune}})}. \quad (9)$$

However, the present application can*not* directly measure $\mathbf{R}_{\mathrm{I+N}}$ but can measure only $\mathbf{R}_{\mathrm{AVS}}$; hence, MVDR beamforming is inapplicable here.

Nonetheless, the beamformer output power may still be minimized by the "minimum-power distortionless-response" (MPDR) beamforming algorithm,[37,38] under a wide class of signal-and-noise statistics. This MPDR beamformer substitutes the unobservable $\mathbf{R}_{\mathrm{I+N}}$ in (10) by the collected data's $\mathbf{R}_{\mathrm{AVS}}$. That is, the MPDR-beamformer weight vector equals

$$\mathbf{w}_{\mathrm{MPDR}}(\theta_{\mathrm{tune}},\phi_{\mathrm{tune}})=\frac{\mathbf{R}_{\mathrm{AVS}}^{-1}\mathbf{a}(\theta_{\mathrm{tune}},\phi_{\mathrm{tune}})}{\mathbf{a}^{T}(\theta_{\mathrm{tune}},\phi_{\mathrm{tune}})\mathbf{R}_{\mathrm{AVS}}^{-1}\mathbf{a}(\theta_{\mathrm{tune}},\phi_{\mathrm{tune}})}. \quad (10)$$

Indeed, MPDR beamforming or MVDR beamforming has been applied to acoustic vector sensors.[14,17,29–31,33,34] The only prior knowledge required in the preceding text is the sensor array's array manifold and the desired source's incident direction, to set $(\theta_{\mathrm{tune}}, \phi_{\mathrm{tune}})=(\theta_{\mathrm{SOI}}, \phi_{\mathrm{SOI}})$; this is the same prior information as for the SMF beamformer of Sec. III.

Figures 3(a) and 3(b) plot the acoustic vector sensor's MPDR beam pattern for exactly the same three-speakers scenario of Figs. 2(a) and 2(b). Similarly, Figs. 3(c) and 3(d) are counterparts to the six-speakers scenario of Figs. 2(c) and 2(d). Comparing the MPDR beam patterns here against the SMF beam pattern earlier in Figs. 2(c) and 2(d): Although the MPDR beam patterns do not necessarily point their peaks exactly at the SOI (as in the case of the SMF beam pattern), the MPDR beam patterns place a null near the interfering speakers. Moreover, even as the acoustic vector sensor's 3×1 array manifold offers only two degrees of freedom, its MPDR beamformer succeeds in Figs. 3(c) and 3(d) to place its null near all five interfering speakers.

The MPDR beamformer array gain may be obtained, by substituting Eq. (10) into Eq. (6) and by applying the "matrix inversion lemma" to $\mathbf{R}_{\mathrm{AVS}}$, to give[39]

$$\begin{aligned}&\mathrm{G}(\mathbf{w}_{\mathrm{MVDR}}(\theta_{\mathrm{tune}},\phi_{\mathrm{tune}}))\\&=\frac{\left|\mathbf{a}^{T}(\theta_{\mathrm{tune}},\phi_{\mathrm{tune}})\mathbf{R}_{\mathrm{AVS}}^{-1}\mathbf{a}(\theta_{\mathrm{SOI}},\phi_{\mathrm{SOI}})\right|^{2}}{\mathbf{a}^{T}(\theta_{\mathrm{tune}},\phi_{\mathrm{tune}})\mathbf{R}_{\mathrm{AVS}}^{-1}\rho_{\mathrm{I+N}}\mathbf{R}_{\mathrm{AVS}}^{-1}\mathbf{a}(\theta_{\mathrm{tune}},\phi_{\mathrm{tune}})}\\&=\frac{B^{2}(\theta_{\mathrm{tune}},\phi_{\mathrm{tune}})}{\frac{(1+\kappa)^{2}}{\mathbf{a}^{T}(\theta_{\mathrm{tune}},\phi_{\mathrm{tune}})\rho_{\mathrm{I+N}}^{-1}\mathbf{a}(\theta_{\mathrm{tune}},\phi_{\mathrm{tune}})}-\frac{B^{2}(\theta_{\mathrm{tune}},\phi_{\mathrm{tune}})(2+\kappa)\kappa}{\mathbf{a}^{T}(\theta_{\mathrm{SOI}},\phi_{\mathrm{SOI}})\rho_{\mathrm{I+N}}^{-1}\mathbf{a}(\theta_{\mathrm{SOI}},\phi_{\mathrm{SOI}})}},\end{aligned} \quad (11)$$

where

$$\kappa=\mathrm{SINR}_{\mathrm{in}}\ \mathbf{a}^{T}(\theta_{\mathrm{SOI}},\phi_{\mathrm{SOI}})\rho_{\mathrm{I+N}}^{-1}\mathbf{a}(\theta_{\mathrm{SOI}},\phi_{\mathrm{SOI}})$$

refers to the $\mathrm{SINR}_{\mathrm{out}}$ of an hypothetical MVDR beamformer steered toward $(\theta_{\mathrm{SOI}}, \phi_{\mathrm{SOI}})$. This MVDR beamformer uses $\mathbf{R}_{\mathrm{I+N}}$ instead of the MPDR beamformer's $\mathbf{R}_{\mathrm{AVS}}$ in Eq. (10). Moreover,

$$B(\theta_{\mathrm{tune}},\phi_{\mathrm{tune}})=\frac{\mathbf{a}^{T}(\theta_{\mathrm{tune}},\phi_{\mathrm{tune}})\rho_{\mathrm{I+N}}^{-1}\mathbf{a}(\theta_{\mathrm{SOI}},\phi_{\mathrm{SOI}})}{\mathbf{a}^{T}(\theta_{\mathrm{tune}},\phi_{\mathrm{tune}})\rho_{\mathrm{I+N}}^{-1}\mathbf{a}(\theta_{\mathrm{tune}},\phi_{\mathrm{tune}})} \quad (12)$$

represents the MVDR beamformer's beam pattern at $(\theta_{\mathrm{tune}}, \phi_{\mathrm{tune}})$. Unlike a conventional array of displaced microphones, the acoustic vector sensor's array gain is frequency *independent* because the acoustic vector sensor's array manifold itself is frequency independent.

For the three-speaker scenario in Fig. 1 (i.e., the SOI and interfering speakers 1 and 5) at $\mathrm{INR}=10$ dB, the preceding $\mathrm{G}(\mathbf{w}_{\mathrm{MPDR}}(\theta_{\mathrm{SOI}}, \phi_{\mathrm{SOI}}))$ attains a 6.7 dB of array gain. This compares favorably with the 3.7 dB achievable by the SMF beamformer.

V. IF PERFECT FOREKNOWLEDGE OF $(\theta_{\mathrm{SOI}}, \phi_{\mathrm{SOI}})$ IS UNAVAILABLE

What if perfect foreknowledge of $(\theta_{\mathrm{SOI}}, \phi_{\mathrm{SOI}})$ is unavailable? Section V A will discuss the estimation of an active speaker's DOA using the acoustic vector sensor at the

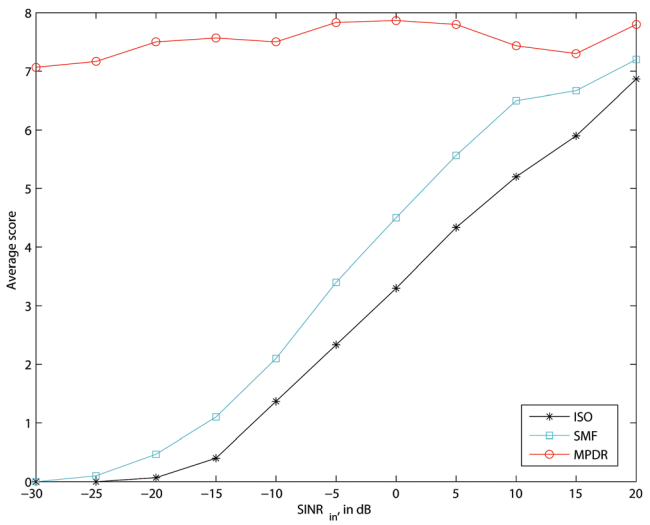

FIG. 6. (Color online) The jury's average score versus $\mathrm{SINR}_{\mathrm{in}}$, for the three-speaker scenario *without* pointing error.

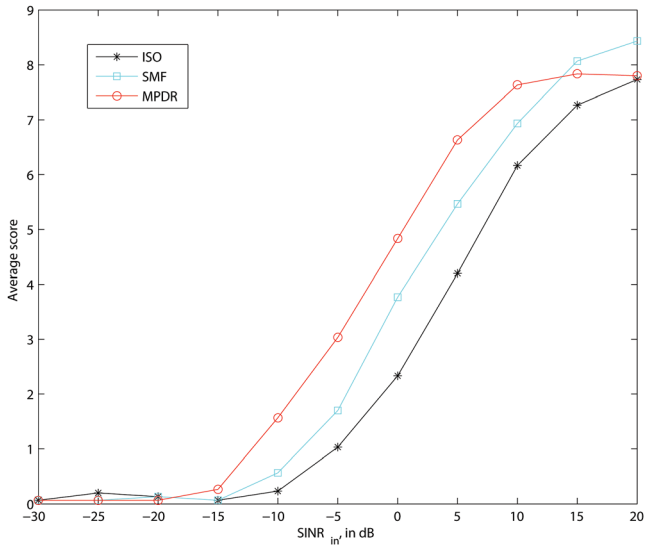

FIG. 7. (Color online) The jury's average score versus $\mathrm{SINR}_{\mathrm{in}}$, for the six-speaker scenario *without* pointing error.

TABLE II. The 15 jurists' personal scores for these $I = 2$ scenarios with perfect beamformer pointing: scenario a, the isotropic microphone (ISO); scenario c, an AVS with the SMF beamformer; and scenario f, an AVS with the MPDR beamformer. The score ranges from 0 for the worst intelligibility to 10 for the best.

Three-speaker scenario without pointing error			
$SINR_{in}$ (dB)	ISO	AVS SMF beamformer	AVS MPDR beamformer
−30	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0	8.0 6.0 7.0 7.0 6.0
	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0	9.0 6.5 6.0 7.0 7.0
	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0	6.0 6.0 9.0 8.0 7.5
	average = 0.0	average = 0.0	average = 7.1
−25	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0	8.0 7.0 7.0 7.0 7.0
	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0	9.0 6.5 6.0 8.0 7.0
	0.0 0.0 0.0 0.0 0.0	0.5 0.0 0.0 0.0 0.0	6.0 6.0 9.0 7.5 6.5
	average = 0.0	average = 0.1	average = 7.2
−20	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0	8.0 7.5 8.0 7.0 6.0
	0.0 0.0 0.0 0.0 0.0	2.0 0.0 0.0 2.0 0.0	9.0 7.0 6.0 8.0 9.0
	0.0 0.0 1.0 0.0 0.0	0.0 0.0 2.0 1.0 0.0	7.0 6.5 9.0 8.0 6.5
	average = 0.1	average = 0.5	average = 7.5
−15	0.0 0.0 0.0 0.0 0.0	0.0 0.0 1.0 0.0 1.0	8.0 7.5 8.0 7.5 6.0
	1.0 0.0 1.0 0.0 0.0	1.0 2.0 1.0 0.0 2.0	9.0 7.5 6.0 9.0 7.0
	0.0 1.5 1.5 1.0 0.0	0.0 3.0 2.5 1.5 1.5	8.0 7.0 9.0 7.0 7.0
	average = 0.4	average = 1.1	average = 7.6
−10	0.5 0.0 1.0 0.0 1.0	1.0 0.0 2.0 0.0 1.0	8.0 7.5 8.0 8.0 7.0
	3.0 3.0 2.0 3.0 1.0	3.0 3.5 2.0 3.0 2.0	9.0 7.5 6.5 8.0 4.0
	0.0 2.0 2.0 1.5 0.5	2.0 4.0 4.0 3.0 1.0	8.0 7.5 9.0 8.0 6.5
	average = 1.4	average = 2.1	average = 7.5
−5	1.0 1.0 2.0 1.0 1.0	2.0 1.0 4.0 2.0 1.0	8.5 7.5 8.5 8.0 7.0
	3.0 4.0 3.0 4.0 2.0	4.0 5.0 5.0 4.0 3.0	9.0 8.0 6.5 9.0 6.5
	0.0 3.0 5.0 4.0 1.0	2.0 5.0 7.0 4.0 2.0	8.0 7.0 9.0 8.0 7.0
	average = 2.3	average = 3.4	average = 7.8
0	2.0 3.0 3.0 3.0 2.0	4.0 4.0 4.0 3.0 3.0	8.5 8.0 8.5 8.0 7.0
	3.0 4.5 5.0 3.02.0	3.0 5.5 6.0 5.0 4.0	9.0 8.0 7.0 8.0 7.0
	3.0 3.0 7.0 5.0 1.0	5.0 6.0 7.5 5.0 2.5	8.0 7.5 9.0 8.0 6.5
	average = 3.3	average = 4.5	average = 7.9
5	4.5 4.0 4.0 4.0 3.0	6.0 6.0 6.0 4.0 4.0	9.0 8.0 8.5 7.0 8.0
	3.0 5.0 5.5 6.0 3.0	5.0 6.0 6.0 7.0 5.0	9.0 7.5 7.5 7.0 7.0
	3.0 4.0 7.5 6.5 2.0	6.0 5.0 7.5 7.0 3.0	8.0 7.5 9.0 8.0 6.0
	average = 4.3	average = 5.6	average = 7.8
10	6.5 6.0 4.0 4.0 4.0	7.5 7.0 6.5 7.0 6.0	9.0 8.0 8.0 6.0 7.0
	2.5 5.5 5.5 7.0 6.0	5.0 6.0 6.0 8.0 6.0	9.0 6.0 7.0 8.0 6.0
	3.0 4.5 7.5 7.0 5.0	6.0 5.0 8.0 7.5 6.0	8.0 6.5 9.0 7.5 6.5
	average = 5.2	average = 6.5	average = 7.4
15	7.5 7.0 5.0 5.0 5.0	8.0 7.5 6.5 6.0 6.0	9.0 7.5 8.0 4.0 7.0
	4.0 6.0 6.0 7.0 6.0	5.0 6.5 6.0 9.0 6.0	9.0 6.5 7.0 8.0 7.0
	4.0 5.0 8.0 7.5 5.5	6.0 6.0 8.5 8.0 5.0	7.0 6.5 9.0 8.0 6.0
	average = 5.9	average = 6.7	average = 7.3
20	8.5 8.0 6.0 5.5 7.0	9.0 8.0 7.5 7.5 7.0	9.5 7.0 8.0 7.0 8.0
	7.0 6.5 7.0 7.0 8.0	8.0 7.0 7.0 8.0 2.0	9.0 6.5 7.0 8.0 7.0
	5.0 5.0 8.5 8.0 6.0000	7.0 5.5 8.5 9.0 7.0000	9.0 7.0 9.0 9.0 6.0000
	average = 6.9	average = 7.2	average = 7.8

receiver. Section V B will discuss the effects on the beamformer where pointing errors exist, i.e., where $(\theta_{tune}, \phi_{tune})$ is only approximately $(\theta_{SOI}, \phi_{SOI})$.

A. To estimate $(\theta_{SOI}, \phi_{SOI})$

If the user cannot manually tune $(\theta_{tune}, \phi_{tune})$ to $(\theta_{SOI}, \phi_{SOI})$, the tuning may be performed electronically and "blindly." Here, blindness refers to the unavailability of any prior knowledge of $(\theta_{SOI}, \phi_{SOI})$. one such estimation method is the "MUltiple SIgnal Classification" (MUSIC):[40,41]

$$(\hat{\theta}_{SOI}, \hat{\phi}_{SOI}) = \arg\max_{\theta,\phi} \frac{1}{\mathbf{a}^T(\theta,\phi)\mathbf{U}\mathbf{U}^T\mathbf{a}(\theta,\phi)}, \quad (13)$$

where $\mathbf{U}$ is a matrix the columns of which are the eigenvectors corresponding to the smallest two eigenvalues of $\mathbf{R}_{AVS}$.

Figure 4 shows the estimation bias $\left(\left|\hat{\theta}_{SOI} - \theta_{SOI}\right|, \left|\hat{\phi}_{SOI} - \phi_{SOI}\right|\right)$ thus obtainable. Each icon in Fig. 4 is based on 200 independent Monte Carlo trials using a 11.3 s speech segment, time-sampled at 44.1 kHz to produce roughly 500 000 time samples to construct $\mathbf{R}_{AVS}$. At INR = 20 dB and

TABLE III. The 15 jurists' personal scores for these $I = 5$ scenarios with perfect beamformer pointing: scenario a, the isotropic microphone (ISO); scenario c, an AVS with the SMF beamformer; and scenario f, an AVS with the MPDR beamformer. The score ranges from 0 for the worst intelligibility to 10 for the best.

Six-speaker scenario without pointing error			
$SINR_{in}$ (dB)	ISO	AVS SMF beamformer	AVS MPDR beamformer
−30	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0
	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0
	0.0 0.0 0.0 1.0 0.0	0.0 0.0 0.0 1.0 0.0	0.0 0.0 0.0 1.0 0.0
	average = 0.1	average = 0.1	average = 0.1
−25	0.0 0.0 0.0 1.0 1.0	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0
	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0
	0.0 0.0 0.0 1.0 0.0	0.0 0.0 0.0 1.0 0.0	0.0 0.0 0.0 1.0 0.0
	average = 0.2	average = 0.1	average = 0.1
−20	0.0 0.0 0.0 1.0 0.0	0.0 0.0 1.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0
	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0
	0.0 0.0 0.0 1.0 0.0	0.0 0.0 0.0 1.0 0.0	0.0 0.0 0.0 1.0 0.0
	average = 0.1	average = 0.1	average = 0.1
−15	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0
	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0
	0.0 0.0 0.0 1.0 0.0	0.0 0.0 0.0 1.0 0.0	0.0 0.0 3.0 1.0 0.0
	average = 0.1	average = 0.1	average = 0.3
−10	0.0 0.0 0.0 0.0 0.0	0.0 0.0 1.0 0.0 0.0	0.0 0.0 1.0 0.0 1.0
	0.0 0.0 1.0 0.0 0.0	2.0 1.0 0.0 0.0 0.0	5.0 3.0 1.0 0.0 0.0
	0.0 0.5 0.0 2.0 0.0	0.0 0.5 1.0 3.0 0.0	0.0 1.0 6.0 4.0 1.5
	average = 0.2	average = 0.6	average = 1.6
−5	0.0 0.0 1.0 1.0 0.0	0.5 0.0 2.0 0.0 1.0	1.5 1.0 2.0 0.0 1.0
	4.0 3.0 1.0 0.0 0.0	3.0 3.0 3.0 0.5 0.0	6.0 6.0 3.0 1.0 3.0
	0.0 0.5 2.0 3.0 0.0	0.0 1.5 4.0 6.0 1.0	1.0 3.0 7.0 8.5 1.5
	average = 1.0	average = 1.7	average = 3.0
0	0.5 1.0 2.0 0.0 1.0	1.0 1.0 4.0 1.0 2.0	1.5 2.0 4.0 0.0 3.0
	4.0 4.5 3.0 0.5 2.0	6.5 7.0 5.0 1.0 4.0	7.5 8.0 6.0 3.0 6.5
	0.5 2.0 5.0 8.0 1.0	3.0 4.0 7.0 8.5 1.5	7.0 5.0 8.0 9.0 2.0
	average = 2.3	average = 3.8	average = 4.8
5	1.5 2.0 5.0 2.0 2.0	4.0 3.0 6.0 3.0 3.0	5.5 4.0 7.0 7.0 4.0
	7.0 5.5 6.0 2.0 4.5	7.5 7.5 6.0 4.0 6.0	8.0 8.5 7.0 6.0 6.5
	4.0 4.0 7.0 9.0 1.5	7.0 6.0 8.0 9.0 2.0	9.0 6.5 8.5 9.0 3.0
	average = 4.2	average = 5.5	average = 6.6
10	5.5 4.0 7.0 5.0 3.0	7.5 6.0 8.0 6.0 4.0	8.0 6.0 8.0 8.0 6.0
	8.0 7.5 7.0 7.0 6.5	8.0 8.0 7.0 7.0 7.0	8.5 8.5 7.5 8.0 7.0
	8.0 6.0 7.0 9.0 2.0	8.0 6.5 8.0 9.0 4.0	8.0 6.5 8.5 9.0 7.0
	average = 6.2	average = 6.9	average = 7.6
15	8.0 6.0 8.0 7.0 5.0	9.5 7.0 8.0 7.0 7.0	10.0 7.0 8.0 6.0 6.0
	8.0 8.0 8.0 7.0 7.0	8.5 8.5 8.0 7.0 9.0	9.0 8.5 7.0 8.5 9.0
	8.0 6.0 8.0 9.0 6.0	10.0 6.5 8.5 9.5 7.0	8.0 6.0 9.0 8.5 7.0
	average = 7.3	average = 8.1	average = 7.8
20	9.5 7.0 9.0 6.0 5.0	10.0 8.0 9.0 8.0 6.0	10.0 7.0 8.0 7.0 6.0
	8.0 8.0 8.0 8.0 9.0	8.5 8.5 8.0 9.0 10.0	9.0 8.5 6.5 7.0 8.0
	8.0 6.0 8.5 9.0 7.0	8.0 7.0 9.0 9.5 8.0	8.0 7.0 9.5 9.5 6.0
	average = 7.7	average = 8.4	average = 7.8

$SINR_{in} \geq 17$ dB, $\hat{\theta}_{SOI}$ and $\hat{\phi}_{SOI}$ are shown there to have such small biases to be under 1°. In a conferencing scenario, social etiquette renders it highly unlikely that more than a couple of conferees would be talking simultaneously. Hence, these simulations assume $I = 2$, i.e., three simultaneous speakers.

B. Beamforming with "look direction" error, where $(\theta_{tune}, \phi_{tune}) \neq (\theta_{SOI}, \phi_{SOI})$

Although the acoustic vector sensor can be manually tuned by the user to point toward the desired speaker, pointing error may occur such that $(\theta_{tune}, \phi_{tune}) \neq (\theta_{SOI}, \phi_{SOI})$. This would degrade the beamformer's performance because the beamformer may regard the SOI as interference and would try to null it. There the array-gain degrades significantly for even a small pointing error, when INR < SNR (i.e., $SINR_{in}$ exceeds roughly 0 dB). However, the array gain is robust to pointing errors, when INR ≥ SNR (i.e., $SINR_{in}$ is under roughly 0 dB).

"Diagonal loading"[42,43] is widely used to mitigate against possible point error without reducing the beamformer's degree of freedom. The diagonally loaded beamforming weight vector equals

$$\mathbf{w}_{\text{MPDR-DL}} = \frac{(\mathbf{R}_{\text{AVS}} + \mathcal{P}_\ell \mathbf{I})^{-1}\mathbf{a}(\theta_{\text{tune}}, \phi_{\text{tune}})}{\mathbf{a}^T(\theta_{\text{tune}}, \phi_{\text{tune}})(\mathbf{R}_{\text{AVS}} + \mathcal{P}_\ell \mathbf{I})^{-1}\mathbf{a}(\theta_{\text{tune}}, \phi_{\text{tune}})}, \quad (14)$$

where $\mathbf{I}$ denotes a 3×3 identity matrix. Diagonal loading thus adds $\mathcal{P}_\ell$ to each diagonal element of $\mathbf{R}_{\text{AVS}}$.

Figure 5(a) and 5(b) show the efficacy of "diagonal loading" to mitigate beamforming pointing error. In Fig. 5(a), where no diagonal loading is applied, the MPDR-DL beamformer is equivalent to the MPDR beamformer. It mistakenly places nulls near the SOI. In Fig. 5(b), where the diagonal loading of $\mathcal{P}_\ell = 316.2$ is applied, the MPDR-DL beamformer successfully places nulls near the interferences but not near the SOI.

VI. JURY TESTS ON THE PROPOSED SCHEMES' EFFECTIVENESS IN SPEECH ENHANCEMENT

The aforementioned reception methods—the isotropic sensor, the SMF beamformer, and the MPDR beamformer—have their output-signals compared here subjectively by a human jury. The jury consists of 15 native speakers of Mandarin Chinese, 4 female and 11 male, aged 23–34. Each jurist, after listening to a speech sample, assigns a score (0 = worst, 10 = best) based on his/her personal perception of that speech-sample's speech intelligibility. Scores ≤ 3 would mean no intelligibility. Scores ≥ 7 would refer to various degrees of speech quality, all with total intelligibility.

The set of jury-tested speech samples cover these six reception scenarios:

(a) the single isotropic sensor (ISO) of Sec. II A.
(b) an acoustic vector sensor with the SMF beamforming of Sec. III at the perfect pointing direction.
(c) an acoustic vector sensor with the MPDR beamformer of Sec. IV at the perfect pointing direction.
(d) an acoustic vector sensor with the SMF beamformer of Sec. III but subject to pointing error.
(e) an acoustic vector sensor with the MPDR beamformer of Sec. IV but subject to pointing error.
(f) an acoustic vector sensor with the MPDR-DL beamformer while subjected to pointing error—see Sec. V B.

For each preceding scenario, there exist two sub-scenarios: with $I = 2$ (involving interfering speakers 1 and 5 of Fig. 1), with $I = 5$. Hence, there exist altogether 12 sub-scenarios. For each of these 12 sub-scenarios, test samples are prepared at various SINR_{in}, but all at $\text{SNR} = 20\,\text{dB}$, $\mathcal{P}_\ell = 10$, and a pointing-error of $\theta_{\text{tune}} - \theta_{\text{SOI}} = 15^\circ$ and $\phi_{\text{tune}} - \phi_{\text{SOI}} = 15^\circ$.

A. Scenarios without beamformer pointing error

Figures 6 and 7 plot the jury's average score versus the SINR_{in}, for $I = 2$ and $I = 5$, respectively. The jurists' corresponding personal scores are listed in Tables II and III. From these figures and tables:

(1) The SMF beamformer improves over the ISO case by only 0 to 1.5 points. The MPDR beamformer improves by 0 to 7 points.
(2) At $I = 2$, the MPDR beamformer offers intelligible speech even for the most adverse SNR of $-30\,\text{dB}$, whereas the SMF beamformer requires an SNR of 18 dB and the ISO requires >20 dB. Here, the number $(I + 1)$ of active speakers does not exceed the acoustic vector sensor's degree of freedom (namely 3); hence, the MPDR beamformer can thus null both interferers while passing the desired speaker to improve speech intelligibility.
(3) At $I = 5$, the MPDR beamformer offers intelligible speech for $\text{SNR} \geq 5\,\text{dB}$, whereas the SMF beamformer would require an $\text{SNR} \geq 10$ dB and the ISO case needs $\text{SNR} \geq 15\,\text{dB}$. Here, the active speakers are more numerous than the acoustic vector sensor's degree of freedom; and the MPDR beamformer is overwhelmed.

B. Scenarios with beamformer "look direction" error

Figures 8 and 9 plot the jury's average score versus the SINR_{in}, for $I = 2$ and $I = 5$ respectively, with the beamformer "look direction" error at the sizeable value of $\theta_{\text{tune}} - \theta_{\text{SOI}} = 15^\circ$,

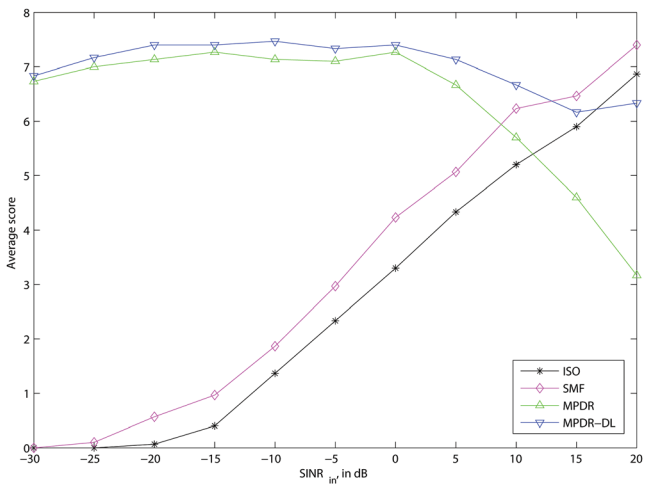

FIG. 8. (Color online) The jury's average score versus SINR_{in}, for the three-speaker scenario with pointing error of $\theta_{\text{tune}} - \theta_{\text{SOI}} = 15^\circ$, $\phi_{\text{tune}} - \phi_{\text{SOI}} = 15^\circ$.

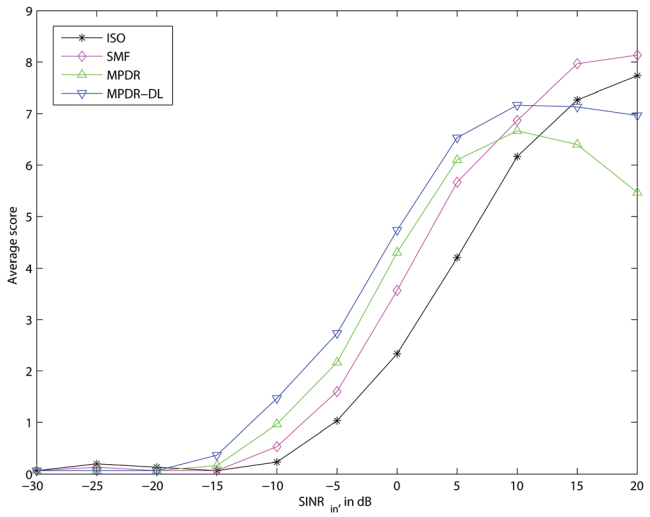

FIG. 9. (Color online) The jury's average score versus SINR_{in}, for the six-speaker scenario with pointing error of $\theta_{\text{tune}} - \theta_{\text{SOI}} = 15^\circ$, $\phi_{\text{tune}} - \phi_{\text{SOI}} = 15^\circ$.

$\phi_{\text{tune}} - \phi_{\text{SOI}} = 15°$. The jurists' corresponding personal scores are listed in Tables IV and V. From these figures and tables:

(1) The SMF beamformer improves over the ISO case by only 0 to 1 point. The MPDR-DL beamformer improves by −0.5 to 7 points.
(2) At $I = 2$, the MPDR-DL beamformer offers intelligible speech even for the most adverse SNR of −30 dB. Here, the number $(I + 1)$ of active speakers does not exceed the acoustic vector sensor's degree of freedom (namely 3); and the MPDR-DL beamformer can thus null both interferers while passing the desired speaker to improve speech intelligibility. However, at high SNR of 30 dB, the MPDR-DL beamformer slightly compromises intelligibility by about half a point.
(3) At $I = 5$, the MPDR beamformer offers intelligible speech for SNR ≥ 5 dB, whereas the SMF beamformer would require an SNR ≥ 10 dB and the ISO case needs SNR ≥ 15 dB. Here, the active speakers are more numerous than the acoustic vector sensor's degree of freedom; and the MPDR beamformer is overwhelmed.

TABLE IV. The 15 jurists' personal scores for these $I = 2$ scenarios subject to pointing error: scenario b, an AVS with the SMF beamformer; scenario d, an AVS with the MPDR beamformer; and scenario e, an AVS with the MPDR-DL beamformer. The score ranges from 0 for the worst intelligibility to 10 for the best.

Three-speaker scenario with pointing error			
SINR_{in} (dB)	AVS SMF beamformer	AVS MPDR beamformer	AVS MPDR-DL beamformer
−30	0.0 0.0 0.0 0.0 0.0	8.0 6.0 7.0 7.0 6.0	8.0 6.0 7.0 7.0 6.0
	0.0 0.0 0.0 0.0 0.0	8.5 6.5 6.0 4.0 6.0	9.0 6.5 6.0 5.0 6.0
	0.0 0.0 0.0 0.0 0.0	5.0 6.0 9.0 9.0 7.0	5.0 6.0 9.0 9.0 7.0
	average = 0.0	average = 6.7	average = 6.8
−25	0.0 0.0 0.0 0.0 0.0	8.0 7.0 7.0 7.0 6.0	8.0 7.0 7.0 7.0 7.0
	0.5 0.0 0.0 0.0 0.0	9.0 6.5 6.0 7.0 7.0	9.0 7.0 6.0 8.0 7.0
	0.0 0.0 1.0 0.0 0.0	5.0 6.5 9.0 8.0 6.0	5.0 6.0 9.0 8.0 6.5
	average = 0.1	average = 7.0	average = 7.2
−20	0.0 0.0 0.0 0.0 0.0	8.0 7.5 8.0 7.0 6.0	8.0 7.5 8.0 7.0 6.0
	1.5 0.0 0.0 3.0 0.0	9.0 6.5 6.0 7.0 7.0	9.0 7.0 6.0 8.0 8.0
	0.0 1.0 2.0 1.0 0.0	5.0 7.0 9.0 8.0 6.0	6.0 7.0 9.0 8.0 6.5
	average = 0.6	average = 7.1	average = 7.4
−15	0.5 0.0 1.0 0.0 1.0	8.0 7.5 8.0 7.0 6.0	8.0 7.5 8.0 7.0 6.0
	1.0 2.0 1.0 0.0 1.0	9.0 7.0 6.0 7.0 6.0	9.0 7.5 6.0 8.0 7.0
	0.0 2.0 2.5 1.5 1.0	7.0 7.0 8.5 8.0 7.0	7.0 7.5 9.0 7.0 6.5
	average = 1.0	average = 7.3	average = 7.4
−10	1.5 1.0 1.0 0.0 1.0	8.0 7.5 8.0 6.0 6.0	8.0 7.5 8.0 7.5 7.0
	3.5 3.0 2.0 2.0 2.0	9.0 7.5 6.5 6.0 6.0	9.0 7.5 6.5 8.0 6.0
	0.0 3.0 4.0 3.0 1.0	7.0 7.0 8.5 8.0 6.0	8.0 7.0 8.5 7.0 6.5
	average = 1.9	average = 7.1	average = 7.5
−5	1.5 1.0 3.0 1.0 1.0	8.5 7.5 8.0 7.0 7.0	8.5 7.5 8.0 7.5 7.0
	4.0 5.0 4.0 3.0 4.0	9.0 7.5 6.5 6.0 3.0	9.0 7.5 6.5 7.0 5.0
	0.0 4.0 7.0 4.0 2.0	7.0 7.0 8.5 7.5 6.5	7.0 6.5 8.5 7.5 7.0
	average = 3.0	average = 7.1	average = 7.3
0	4.0 4.0 4.0 3.0 3.0	8.5 7.5 7.5 7.0 7.0	8.5 7.5 7.5 7.0 7.0
	3.0 5.0 6.0 4.0 4.0	8.0 7.5 6.5 7.0 6.0	8.5 7.5 6.5 8.0 6.0
	5.0 4.0 7.5 5.0 2.0	7.0 6.5 8.5 7.5 7.0	7.0 7.0 8.5 7.5 7.0
	average = 4.2	average = 7.3	average = 7.4
5	6.0 6.0 6.0 4.0 3.0	8.5 7.5 7.0 6.0 6.0	8.5 7.5 7.5 6.5 7.0
	4.0 5.0 6.0 7.0 5.0	7.0 7.5 6.5 6.0 6.0	7.5 7.5 7.0 7.0 6.5
	3.0 4.0 7.5 7.0 2.5	7.0 6.0 7.5 7.5 4.0	7.0 7.0 8.0 7.5 5.0
	average = 5.1	average = 6.7	average = 7.1
10	7.0 8.0 6.0 7.0 5.0	7.5 7.0 5.0 4.0 5.0	8.0 7.0 6.5 5.0 6.0
	4.0 5.5 6.0 8.0 6.0	5.0 5.0 6.0 7.0 6.0	8.0 5.5 6.5 7.0 6.0
	5.0 5.0 8.0 7.5 5.5	4.0 5.5 7.0 6.0 5.5	7.0 6.0 8.0 7.5 6.0
	average = 6.2	average = 5.7	average = 6.7
15	7.5 8.0 6.0 6.0 6.0	5.5 6.0 4.0 2.0 4.0	8.5 7.0 6.0 3.0 5.0
	4.0 6.0 6.0 8.0 6.0	6.0 4.0 5.0 6.0 6.0	8.5 5.0 6.0 7.0 7.0
	5.0 6.0 8.5 8.0 6.0	3.0 4.5 5.0 5.0 3.0	4.0 5.0 8.0 7.5 5.0
	average = 6.5	average = 4.6	average = 6.2
20	9.0 8.5 7.5 6.5 7.0	1.0 5.0 3.0 0.5 3.0	8.0 6.0 6.5 6.0 6.0
	8.0 6.5 6.0 8.0 8.0	2.0 3.0 4.5 5.0 4.0	7.5 5.0 6.0 8.0 4.0
	6.0 6.0 8.5 8.5 7.0	3.0 4.0 3.0 4.0 2.5	6.0 6.0 8.0 7.5 4.5
	average = 7.4	average = 3.2	average = 6.3

The MPDR beamformer outperforms the ISO and the SMF beamformer when $SINR_{in}$ is smaller than about 5 dB. However, as $SINR_{in}$ is larger than 5 dB, the performance of the MPDR beamformer largely degrades and the average score drops below the SMF beamformer and the ISO. This is because the MPDR beamformer treats the SOI as an interference for the pointing error scenario, and this SOI canceling becomes significant when $SINR_{in}$ is large. The MPDR-DL beamformer, in contrast, significantly compensates the SOI canceling effect. At $SINR_{in} = 20$ dB, the average score of MPDR-DL beamformer is about 3.1 and 1.5 higher than the MPDR beamformer for three speakers and six speakers scenario, respectively. Thus, the MPDR-DL beamformer balances the performance in the low and high $SINR_{in}$ ranges. At low $SINR_{in}$, the MPDR-DL beamformer has a good performance as the MPDR beamformer does. At high $SINR_{in}$, the MPDR-DL beamformer does not suffer the severe SOI canceling as the MPDR beamformer suffers. Hence, MPDR-DL beamformer is more robust than the MPDR beamformer against the pointing error. The average score of the MPDR-DL beamformer drops

TABLE V. The 15 jurists' personal scores for these $I = 5$ scenarios subject to pointing error: scenario b, an AVS with the SMF beamformer; scenario d, an AVS with the MPDR beamformer; and scenario e, an AVS with the MPDR-DL beamformer. The score ranges from 0 for the worst intelligibility to 10 for the best.

Six-speaker scenario with pointing error			
$SINR_{in}$ (dB)	AVS SMF beamformer	AVS MPDR beamformer	AVS MPDR-DL beamformer
−30	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0
	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0
	0.0 0.0 0.0 1.0 0.0	0.0 0.0 0.0 1.0 0.0	0.0 0.0 0.0 1.0 0.0
	average = 0.1	average = 0.1	average = 0.1
−25	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0
	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0
	0.0 0.0 0.0 1.0 0.0	0.0 0.0 0.0 1.0 0.0	0.0 0.0 0.0 1.0 0.0
	average = 0.1	average = 0.1	average = 0.1
−20	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0
	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0
	0.0 0.0 0.0 1.0 0.0	0.0 0.0 0.0 1.0 0.0	0.0 0.0 0.0 1.0 0.0
	average = 0.1	average = 0.1	average = 0.1
−15	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0
	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0	0.0 0.0 0.0 0.0 0.0
	0.0 0.0 0.0 1.0 0.0	0.0 0.5 1.0 1.0 0.0	0.0 0.5 2.0 3.0 0.0
	average = 0.1	average = 0.1	average = 0.4
−10	0.0 0.0 0.0 0.0 0.0	0.0 0.0 1.0 0.0 0.0	0.0 0.0 1.0 0.0 1.0
	1.0 0.5 1.0 0.0 0.0	2.0 2.0 1.0 0.0 0.0	5.0 2.0 1.0 0.0 0.0
	0.0 0.5 2.0 3.0 0.0	0.0 0.5 5.0 3.0 0.0	0.0 1.0 5.5 4.0 1.5
	average = 0.5	average = 1.0	average = 1.5
−5	0.5 0.0 1.0 0.0 1.0	1.0 1.0 1.0 0.0 1.0	1.0 1.0 3.0 0.0 1.0
	3.0 3.0 3.0 0.5 0.0	4.0 3.5 3.0 0.5 0.0	5.0 4.0 3.0 1.0 2.0
	0.0 1.0 4.0 6.0 1.0	0.0 2.0 7.0 7.0 1.5	1.0 2.5 7.0 8.0 1.5
	average = 1.6	average = 2.2	average = 2.7
0	2.0 2.0 2.0 0.0 2.0	2.0 2.0 4.0 0.0 2.0	1.5 2.0 5.0 0.0 3.0
	6.0 7.0 5.0 0.5 3.5	7.0 8.0 6.0 2.0 4.5	7.5 8.0 6.0 3.0 6.0
	3.0 3.5 7.0 8.5 1.5	5.0 4.0 8.0 8.5 1.5	5.0 5.0 8.0 9.0 2.0
	average = 3.6	average = 4.3	average = 4.7
5	4.0 4.0 6.0 6.0 3.0	4.5 3.0 7.0 6.0 4.0	5.5 4.0 7.0 7.0 4.0
	7.5 7.5 6.0 3.0 6.0	8.0 8.5 7.0 5.0 6.0	7.5 8.5 7.0 6.0 6.5
	7.0 6.0 8.0 9.0 2.0	7.0 5.5 8.5 9.0 2.5	8.0 6.5 8.5 9.0 3.0
	average = 5.7	average = 6.1	average = 6.5
10	7.0 6.0 8.0 8.0 4.0	7.5 5.0 7.0 5.0 5.0	7.5 6.0 7.0 6.0 5.0
	8.0 7.5 7.0 7.0 6.5	8.0 8.5 7.0 6.0 7.0	8.5 8.5 7.5 7.0 7.0
	8.0 6.0 8.0 9.0 3.0	8.0 6.0 6.0 9.0 5.0	8.0 6.0 7.5 9.0 7.0
	average = 6.9	average = 6.7	average = 7.2
15	9.5 7.0 8.0 8.0 6.0	8.5 5.0 7.0 3.0 4.0	9.0 6.0 7.0 5.0 5.0
	8.5 8.0 8.0 8.0 8.0	8.5 7.5 7.5 7.0 8.0	8.5 7.5 7.0 7.5 9.0
	9.0 6.5 8.5 9.5 7.0	7.0 6.0 3.0 8.0 6.0	7.0 6.0 7.0 8.5 7.0
	average = 8.0	average = 6.4	average = 7.1
20	10.0 8.0 9.0 7.0 5.0	5.0 6.0 6.0 2.0 3.0	9.5 7.0 7.0 7.0 5.0
	8.5 8.5 8.0 8.5 9.0	3.0 4.0 6.0 9.5 8.0	7.5 7.5 6.5 5.0 7.0
	8.0 7.0 9.0 9.5 7.0	6.0 6.0 2.0 7.5 8.0	7.0 6.0 8.0 8.5 6.0
	average = 8.1	average = 5.5	average = 7.0

below that of the SMF beamformer at the higher $SINR_{in}$, which is the cost of obtaining the robustness. Furthermore, it can be seen that the SMF beamformer is robust against the pointing error, but it has poor performance at the low $SINR_{in}$.

If they had the same (or almost the same) azimuth elevation DOA, then DOA diversity would indeed fail, whether for an acoustic vector sensor or for a linear array of isotropic microphones. The DOA resolution limit has been investigated.[9]

VII. CONCLUSION

The acoustic vector sensor is proposed for the first time in the open literature for speech enhancement. The user tunes the acoustic vector sensor's desired beamforming direction. Only one beamforming weight vector needs be computed for all audio frequencies despite the sound signals' wide bandwidths, time-varying spectra, and temporally nonstationary statistics. No prior knowledge is needed of any aspect of any interference or noise. Jury tests verify the efficacy of the proposed scheme in enhancing speech intelligibility in a conference-room setting involving multiple simultaneous speakers. Also investigated is the case where the acoustic vector sensor must self-tune its beamformer's pointing direction.

ACKNOWLEDGMENTS

The authors were supported by the Internal Competitive Research Grant No. G-YG67 from the Hong Kong Polytechnic University.

[1]Such beamforming may be construed as a spatial counterpart of a passband-stopband frequency filter. Instead of the latter's discrete-time samples, a beamformer collects spatial samples visa multiple sensors distributed over a spatial region.

[2]When any emitter lies in the acoustic vector-sensor's near field, the pressure-sensor's response will be frequency-dependent (Ref. 5). Hence, the present beamformer will omit the pressure sensor to maintain the beamformer's frequency independence for all near-field/far-field sources.

[3]A. Nehorai and E. Paldi, "Acoustic vector-sensor array processing," IEEE Trans. Signal Process. **42**, 2481–2491 (1994).

[4]J. A. McConnell, "Analysis of a compliantly suspended acoustic velocity sensor," J. Acoust. Soc. Am. **113**, 1395–1405 (2003).

[5]Y. I. Wu, K. T. Wong, and S. K. Lau, "The acoustic vector-sensor's near-field array-manifold," IEEE Trans. Signal Process. **58**, 3946-3951 (2010).

[6]B. Hochwald and A. Nehorai, "Identifiability in array processing models with vector-sensor applications," IEEE Trans. Signal Process. **44**, 83–95 (1996).

[7]M. Hawkes and A. Nehorai, "Effects of sensor placement on acoustic vector-sensor array performance," IEEE J. Ocean. Eng. **24**, 33–40 (1999).

[8]B. A. Cray and A. H. Nuttall, "Directivity factors for linear arrays of velocity sensors," J. Acoust. Soc. Am. **110**, 324–331 (2001).

[9]K. T. Wong and H. Chi, "Beam patterns of an underwater acoustic vector hydrophone located away from any reflecting boundary," IEEE J. Ocean. Eng. **27**, 628–637 (2002).

[10]M. T. Silvia and R. T. Richards, "A theoretical and experimental investigation of low-frequency acoustic vector sensors," in *IEEE Oceans Conference* (2002), pp. 1886–1897.

[11]B. A. Cray, "Directional point receivers: The sound and the theory," in *IEEE Oceans Conference* (2002), pp. 1903–1905.

[12]B. R. Rapids and G. C. Lauchle, "Processing of forward scattered acoustic fields with intensity sensors," in *IEEE Oceans Conference* (2002), pp. 1911–1914.

[13]B. A. Cray, V. M. Evora, and A. H. Nuttall, "Highly directional acoustic receivers," J. Acoust. Soc. Am. **113**, 1527–1533 (2003).

[14]S. Guiqing, L. Qihu, and Z. Bin, "Acoustic vector sensor signal processing," Chin. J. Acoust. **25**, 1–14 (2006).

[15]H. Cox, "Super-directivity revisited," in *IEEE Instrumentation and Measurement Technology Conference* (2004), pp. 87–90.

[16]H. Keshavarz, "Beam patterns of an underwater acoustic vector hydrophone located near a reflecting boundary," in *IEEE Oceans Conference* (2004), Vol. 2, pp. 585–588.

[17]G. L. D'Spain, J. C. Luby, G. R. Wilson, and R. A. Gramann, "Vector sensors and vector sensor line arrays: Comments on optimal array gain and detection," J. Acoust. Soc. Am. **120**, 171–185 (2006).

[18]N. Qi and T. Tian, "Acoustic vector hydrophone array supergain energy flux beamforming," in *International Conference on Signal Processing* (2006), Vol. 4.

[19]D. J. Schmidlin, "Directionality of generalized acoustic sensors of arbitrary order," J. Acoust. Soc. Am. **121**, 3569–3578 (2007).

[20]H. Cox and H. Lai, "Performance of line arrays of vector and higher order sensors," in *Asilomar Conference on Signals, Systems and Computers* (2007), pp. 1231–1236.

[21]J. A. Clark, "High-order angular response beamformer for vector sensors," J. Sound Vib. **318**(3), 417–422 (2008).

[22]P. K. Tam and K. T. Wong, "Cramer-rao bounds for direction finding by an acoustic vector sensor under nonideal gain-phase responses, noncollocation, or nonorthogonal orientation," IEEE Sens. J. **9**, 969–982 (2009).

[23]K. T. Wong, "Acoustic vector-sensor ffh blind beamforming & geolocation," IEEE Trans. Aerosp. Electron. Syst. **46**, 444–449 (2010).

[24]Y. I. Wu and K. T. Wong, "Acoustic near-field source localization by two passive anchor nodes," IEEE Trans. Aerosp. Electron. Syst. **48**, 159–169 (2012).

[25]H. Cox and R. M. Zeskind, "Adaptive cardiod processing," in *Asilomar Conference on Signals, Systems and Computers* (1992), pp. 1058–1061.

[26]M. Hawkes and A. Nehorai, "Acoustic vector-sensor beamforming and capon direction estimation," IEEE Trans. Signal Process. **46**, 2291–2304 (1998).

[27]H. Junying, L. Chunxu, L. Guolong, and L. Hong, "A combined signal processing approach against coherent interference with pressure and particle velocity," Chin. J. Acoust. **20**, 2001–2010 (2001).

[28]J. Hui, H. Liu, M. Fan, and G. Liang, "Study on the physical basis of pressure and particle velocity combine processing," Chin. J. Acoust. **20**, 203–212 (2001).

[29]H. W. Chen and J. W. Zhao, "Wideband mvdr beamforming for acoustic vector sensor linear array," IEE Proc., Radar Sonar Navig. **151**, 158–162 (2004).

[30]S. Guiqing and L. Qihu, "Acoustic vector sensor signal processing," Acta Acust. **29**, 491–498 (2004).

[31]M. E. Lockwood and D. L. Jones, "Beamformer performance with acoustic vector sensors in air," J. Acoust. Soc. Am. **119**, 608–619 (2006).

[32]H. Cox and H. Lai, "Endfire supergain with a uniform line array of pressure and velocity sensors," in *Asilomar Conference on Signals, Systems and Computers* (2006), pp. 2271–2275.

[33]H. Lai, H. Cox, and K. Bell, "Adaptive factored beamforming for vector sensor arrays," in *Asilomar Conference on Signals, Systems and Computers* (2008), pp. 402–406.

[34]A. Agarwal, A. Kumar, M. Aggarwal, and R. Bahl, "Design and experimentation with acoustic vector sensors," in *International Symposium on Ocean Electronics* (2009), pp. 139–146.

[35]Please see equation (6.202) on page 491 of Ref. 38.

[36]D. G. Manolakis, *Statistical and Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering and Array Processing*, 641–656 (McGraw-Hill, Boston, MA, 2000).

[37]H. L. V. Trees, *Detection, Estimation, and Modulation Theory, Part IV: Optimum Array Processing* (Wiley, New York, 2002), pp. 439–512.

[38]H. Cox, "Resolving power and sensitivity to mismatch of optimum array processors," J. Acoust. Soc. Am. **54**, 771–785 (1973).

[39]Please see equations (6.218) to (6.230) in Ref. 38.

[40]R. O. Schmidt, "Multiple emitter location and signal parameter estimation," IEEE Trans. Antennas Propag. **AP-34**, 276–280 (1986).

[41]M. Shujau, C. H. Ritz, and I. S. Burnett, "Using in-air acoustic vector sensors for tracking moving speakers," in *International Conference on Signal Processing and Communication Systems* (2010), pp. 1–5.

[42]J. Li, P. Stoica, and Z. Wang, "On robust capon beamforming and diagonal loading," IEEE Trans. Signal Process. **51**, 1702–1715 (2003).

[43]Y. Zhang, D. Sun, and D. Zhang, "Robust adaptive acoustic vector sensor beamforming using automated diagonal loading," Appl. Acoust. **70**, 1029–1033 (2009).