IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 4, 2020, accepted October 9, 2020, date of publication October 13, 2020, date of current version October 26, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3030814

Deconvolved Conventional Beamforming and Adaptive Cubature Kalman Filter Based Distant Speech Perception System

XIANG PAN[1], YUE BAO[1], YITING ZHU[1], HUANGYU DAI[1], AND JIANGFAN ZHANG[2], (Member, IEEE)

[1]School of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China
[2]Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, MO 65401, USA

Corresponding author: Jiangfan Zhang (jiangfanzhang@mst.edu)

The work of Xiang Pan, Yue Bao, and Yiting Zhu was supported in part by the National Natural Science Foundation of China under Grant 41776108, Grant 61571397, and Grant 61171148; and in part by the National Key Research Program of China under Grant 2017YFC0306901. The work of Jiangfan Zhang was supported in part by the Cynthia Tang Endowment in Computer Engineering, Missouri University of Science and Technology, Rolla, MO, USA.

ABSTRACT A spatial-temporal processing framework integrated of speech enhancement and speech tracking is proposed in this paper for distant speech perception. First, weak speech signals are enhanced by the deconvolved conventional beamforming (DCBF) with a microphone array. By virtue of the narrow beamwidth and low sidelobes of the DCBF, the competing sources can be effectively suppressed without introducing extra speech distortion. Second, with the accurate bearing provided by the DCBF, the Cubature Kalman filter can be utilized to track the speech source of interest. By introducing a scaling factor in the current statistical motion model, a new tracking algorithm is proposed which is suitable for both maneuvering and nonmaneuvering speech sources. The introduced scaling factor can be adaptively adjusted to improve the tracking performance of the proposed algorithm for different motion models. Numerical results show that the proposed algorithm can provide better tracking performance than the conventional one. In particular, the tracking root mean square error can be reduced by half for some cases.

INDEX TERMS Cubature Kalman filter, deconvolved conventional beamforming, improved current statistical motion model, maneuvering speech source, speech perception system.

I. INTRODUCTION

Although the automatic speech recognition (ASR) products have been widely implemented in practical applications, most of ASR systems are only suitable for short-range speech source within 5 m. The distant speech perception has not been well studied yet, and is a challenging task due to the severe signal attenuation, interference and background noise [1]–[4]. In indoor environments, reverberation is the main interference [5] while the wind noise is the main interference in outdoor environments [6]. A typical speech enhancement module is proposed in [1], consisting of a speaker tracker, a beamformer, and a post-filter. When the speaker's position is estimated by the tracker, a beamformer is employed to strengthen the sound waves coming from the direction of interest. And the residual noise components are removed by a post-filter. In this paper, we develop a distant speech perception system with novel beamforming algorithms and enhanced maneuvering speech tracking capability.

The minimum-variance distortionless response (MVDR) [7] is widely utilized in speech enhancement, but the MVDR requires multiple snapshots to calculate the sample covariance matrix. For a moving speech, there are no enough snapshots for calculating the covariance matrix. Thus, we consider the deconvolved conventional beamforming (DCBF) to estimate the direction of arrival (DOA) of the speech of interest. The DCBF has some advantages, such as narrow beamwidth, low sidelobes comparable with the MVDR, and moreover, it maintains the robustness of the conventional beamforming (CBF).

It is of great interest to track a moving speaker of interest in practice. The Kalman filter (KF) [8], the extended Kalman filter (EKF) [9], the unscented Kalman filter (UKF) [10], and the particle filter (PF) [11] are widely utilized in

The associate editor coordinating the review of this manuscript and approving it for publication was Yue Zhang.

target tracking. Here, we consider the cubature Kalman filter (CKF) [12] for speech tracking on account of its lower computational complexity than the PF and better stability than the UKF.

Sometimes, the speaker may abruptly stop for a while, and then continue to walk. However, it is difficult to track such a maneuvering target with some unknown time-varying motion model. The current statistical (CS) model [13] can be utilized to model the time-varying motion model. The CS model assumes that given the current acceleration, the probability density function of the acceleration at the next instant is a modified Rayleigh density function whose mean value is the current acceleration. The CS model is essentially an extended Singer model adaptively with a non-zero mean of the acceleration [14]. In [15], the limits of acceleration are adaptively adjusted using the fuzzy control in the presence of strong maneuvering. While in [16], the acceleration variance is adaptively modified by estimation of the positional shift. Besides, the interacting multiple model (IMM) algorithm is widely utilized in tracking a maneuvering target, wherein the target maneuver is modeled as a combination of different motion models, such as a nonzero mean, white noise turn rate dynamic model for tracking sharply maneuvering ground targets [17], a combination of the constant velocity (CV) model and the coordinated turn (CT) model with estimated turn rate [18]. In particular, the IMM algorithm reduces to the autonomous multiple model (AMM) algorithm when the transition probability matrix (TPM) is an identity matrix [19]. In this paper, we propose an adaptive CS model by using a scaling factor to improve the tracking performance of the proposed algorithm for both maneuvering and nonmaneuvering motion models. Different from the previous methods, the scaling factor is directly calculated using the residuals during two successive iterations. And the initial parameters are weighted with the scaling factor for next iteration.

The contributions of the paper include:

(1) A distant speech perception system is proposed by integrating the DCBF algorithm and an improved tracking algorithm. The existing speech perception and tracking techniques are generally only effective for the case where the speech source is at a distance ranging from 3 m to 6 m [20], [21] and require a high signal-to-noise (SNR) at the receiving end. In this paper, we are primarily interested in the case where the speech source is at least 15 m far apart, and the algorithms proposed in this paper perform very well in our outdoor experiments.

(2) We extend the deconvolved conventional beamforming algorithm to a circular array for the DOA estimation and speech enhancement. Moreover, the DCBF proposed in [22] which is only suitable for narrowband signals is extended to be suitable for wideband signals by adopting the incoherent signal-subspace processing scheme in this paper. The developed algorithms have been validated by our numerical and experiment results.

(3) An adaptive CS model is developed to improve the performance of our proposed tracking algorithm for both maneuvering and nonmaneuvering motion models. To be specific, we utilize a time-varying scaling factor to scale the reciprocal of the maneuver time constant in the CS model. It is worth mentioning that once a time-varying scaling factor is introduced to scale the reciprocal of the maneuver time constant in the CS model, the Cubature Kalman filter algorithm, which is the tracking algorithm of our interest in this paper, generally cannot be applied to the CS model anymore. To overcome this, we set the product of the time-varying reciprocal of the maneuver time constant and the sampling period to be small, which is generally the case in practice since the sampling period is generally very small in practical systems. Then, the implementation of the Cubature Kalman filter algorithm is developed based on the proposed improved current statistical model. As demonstrated by the numerical and experimental results, the proposed adaptive CS model works well for both maneuvering and nonmaneuvering speech sources.

The remainder of this paper is organized as follows. In Section II, the DCBF algorithm is derived for wideband speech signals. The adaptive CS model combined with the CKF for speech tracking is introduced in Section III. Section IV numerically evaluates the performance of the DCBF processor and the adaptive CS-based Cubature Kalman filter. Section V briefly describes the hardware platform of the distant speech perception system. Experiment results are presented in Section VI. Section VII summarizes the paper.

II. WIDEBAND DCBF

Since the speech signal generally cover the frequency ranging from 300 Hz to 3400 Hz, we consider wideband beamforming techniques to estimate the DOA of the speech signal. As shown in Fig. 1, the wideband signals are firstly decomposed into a group of narrowband components by using Discrete Fourier Transform (DFT). Then, the narrowband DCBF is employed for estimating the DOA for each frequency bin. Then, the final DOA estimate can be obtained by averaging the DOA estimates over all frequency bins. This idea of extending the DCBF from narrowband signals to the wideband signals is similar to that proposed in [23] which is employed to estimate the spatio-temporal spectrum of the signals received by a passive array. However, the application considered in this paper is different from that in [23]. To be specific, we focus on the deconvolved conventional beamforming approach while the paper [23] considers the eigenstructure methods.

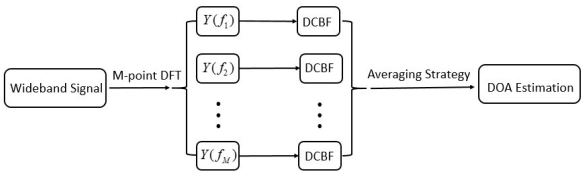


FIGURE 1. The scheme of DOA estimator for wideband signals.

One of major advantages of the circular array is that the shape of its beampattern is invariant to the look direction. A wideband DCBF implemented on a circular array is our focus. Consider an N-element circular array on the x-y plane with radius r and its center at the origin of coordinate.[1] The array elements are evenly distributed on the circumference which spatially sample a sound field at the locations:

$$[x_n, y_n] = \left[r\cos(\frac{2\pi}{N}n), r\sin(\frac{2\pi}{N}n) \right], \quad n = 0, 1, \ldots, N-1. \tag{1}$$

In (1), we only consider the azimuth angle of incident signals. For the case where the signals are not incident on the horizontal plane containing the circular array, the CBF can be firstly used to make a joint estimation of the azimuth and elevation angles. Then, by using the estimated elevation angle, the DCBF is utilized to accurately estimate the azimuth angle.

Assume that there are K speech sources which emit signals to the circular array from far field with different azimuth angles θ_i and a known elevation angle ϕ_0. The received signal can be expressed as

$$X = \sum_{i=1}^{K} A_i p_i + n \tag{2}$$

where A_i is a random signal amplitude and independent over i, n denotes a white Gaussian noise vector with mean 0 and covariance $\sigma^2 I$, which is independent of A_i, and p_i is a source-direction dependent vector, which is defined by

$$\begin{aligned} p_i &= \left[p_1^i, p_2^i, \ldots, p_N^i \right]^T \\ &= \Big[e^{-jk(x_1\cos\phi_0\cos\theta_i + y_1\cos\phi_0\sin\theta_i)}, \\ &\quad e^{-jk(x_2\cos\phi_0\cos\theta_i + y_2\cos\phi_0\sin\theta_i)}, \\ &\quad \ldots, e^{-jk(x_N\cos\phi_0\cos\theta_i + y_N\cos\phi_0\sin\theta_i)} \Big]^T \end{aligned} \tag{3}$$

where j denotes the imaginary unit, k denotes wave number, the superscript T denotes vector transpose. The exponent of p_n^i denotes the phase delay of signal from the center of the circular array to the n-th array element.

In order to estimate the DOA of the signal by the CBF, the steering vector $s = [s_1, s_2, \ldots s_N]^T$ is multiplied by the received signal X, so we can obtain the beam power for different angle, where $s_n = \frac{1}{N} e^{-jk(x_n\cos\phi_0\cos\theta + y_n\cos\phi_0\sin\theta)}$, θ is referred to as the steering or look-direction angle. The expected beam power is given by

$$\begin{aligned} Y(\theta) &= s^H \left[\sum_{i,j=1}^{K} p_i \left\langle A_i A_j^* \right\rangle p_j^H + \left\langle nn^H \right\rangle \right] s \\ &= \sum_{i=1}^{K} \left\langle |A_i|^2 \right\rangle \left| s^H p_i \right|^2 + \sigma^2. \end{aligned} \tag{4}$$

[1]The element and receiver are interchangeable in this paper.

where $<>$ denotes expectation. Note that σ^2 is independent of s and known, and hence we can drop it in (4). As a result,

$$Y(\theta) \propto \sum_{i=1}^{K} \left\langle |A_i|^2 \right\rangle \left| s^H p_i \right|^2 = \int_0^{2\pi} B_p(\theta, \vartheta) S(\vartheta)\, d\vartheta, \tag{5}$$

where $S(\vartheta) \triangleq \sum_{i=1}^{K} \left\langle |A_i|^2 \right\rangle \delta(\vartheta - \theta_i)$ and

$$\begin{aligned} &B_p(\theta, \theta_i) \\ &\triangleq \left| s^H p_i \right|^2 \\ &= \left| \frac{1}{N} \sum_{n=1}^{N} e^{jkr\left[(\cos\theta - \cos\theta_i)\cos\left(\frac{2\pi n}{N}\right) + (\sin\theta - \sin\theta_i)\sin\left(\frac{2\pi n}{N}\right)\right]} \right|^2 \\ &= \left| \frac{1}{N} \sum_{n=1}^{N} e^{j2kr\sin\left(\frac{\theta-\theta_i}{2}\right)\sin\left(\frac{2\pi n}{N} - \frac{\theta+\theta_i}{2}\right)} \right|^2. \end{aligned} \tag{6}$$

Noting that as the number of the array elements $N \to \infty$, (6) can be rewritten as

$$\begin{aligned} B_p(\theta, \theta_i) &= \left| \frac{1}{2\pi} \int_0^{2\pi} e^{j2kr\sin\left(\frac{\theta-\theta_i}{2}\right)\sin\left(\phi - \frac{\theta+\theta_i}{2}\right)} d\phi \right|^2 \\ &= \left| \frac{1}{2\pi} \int_0^{2\pi} e^{j2kr\sin\left(\frac{\theta-\theta_i}{2}\right)\sin(\phi)} d\phi \right|^2 \\ &\triangleq B_p(\theta - \theta_i) \end{aligned} \tag{7}$$

Consequently, (5) can be rewritten as

$$Y(\theta) = \int_0^{2\pi} B_p(\theta - \vartheta) S(\vartheta)\, d\vartheta \tag{8}$$

From (8), we can see that the CBF beam power can be expressed as the convolution of the beam pattern with the source power distribution. Given the beam pattern of an array, we can deconvolve the CBF beam power to estimate the bearing distribution of the sources by using the Richardson-Lucy (RL) algorithm [24], [25], and then we can estimate the DOA of a source of interest. The R-L algorithm of deconvolution is widely used in image processing where the channel impulse response (CIR) is referred to as the point scattering function (PSF) and assumed to be position independent. We apply R-L algorithm to the CBF beam power where the beam pattern is viewed as the PSF to obtain the underlying signal that has been corrupted by the PSF.

III. ADAPTIVE CUBATURE KALMAN FILTER

In this section, an adaptive Kalman filtering algorithm is discussed for maneuvering speech source tracking. The state equation and measurement equation are given by

$$x_l = f(x_{l-1}, u_{l-1}) + n_{l-1} \text{ and } z_l = h(x_l) + w_l. \tag{9}$$

where $x_l = [x_l\ \dot{x}_l\ \ddot{x}_l\ y_l\ \dot{y}_l\ \ddot{y}_l] \in \mathbb{R}^{n_x}$ is the state varible at the time instant l, (x_l, y_l), $(\dot{x}_l, \dot{y}_l)$ and $(\ddot{x}_l, \ddot{y}_l)$ are the position, velocity and acceleration of the speech source, respectively; $f : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_x}$ and $h : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_z}$ are some known functions; $u_l \in \mathbb{R}^{n_u}$ is the control input; $z_l \in \mathbb{R}^{n_z}$ is the measurement; $\{n_{l-1}\}$ and $\{w_l\}$ are independent process

and measurement Gaussian noise sequences with zero means and variances Q_{l-1} and R_l, respectively.

Since the azimuth angle of a speech source is obtained by using beamforming techniques, the measurement equation in (9) can be rewritten as

$$z_l = arctan\left(\frac{y_l - y_o}{x_l - x_o}\right) + w_l \tag{10}$$

where (x_o, y_o) is the position of the observer. It is clear that the measurement equation is nonlinear.

A. IMPROVED MANEUVERING MODEL

The CS model [13] is widely utilized in tracking a maneuvering target. When a target is maneuvering with a certain acceleration, the CS model assumes that its acceleration at the next time instant is limited within a range around the current acceleration. The acceleration can be modeled by a modified Rayleigh density function with mean value equal to the current acceleration. For simplicity, let's first assume that the target moves along one direction, that is, x-axis, and the following results can be easily extended to the case where the target moves arbitrarily in the two-dimensional space. As such, the state vector $x_l = [x_l\ \dot{x}_l\ \ddot{x}_l] \in \mathbb{R}^3$ and the maneuvering model can be expressed as

$$\ddot{x}_l = \bar{a}_l + a_l \quad \text{and } \dot{a}_l = -\alpha a_l + w_l, \tag{11}$$

where $\bar{a}_l$ is the mean of maneuvering acceleration at time instant l. a_l is the zero mean colored acceleration noise. α is the reciprocal of the maneuver time constant, and w_l is white noise with zero mean and variance $\sigma_w^2 = 2\alpha\sigma^2$ [13].

When $\ddot{x}_l > 0$, the probability density function is given by

$$P_r(\ddot{x}_l) = \begin{cases} \dfrac{(a_{max} - \ddot{x}_l)}{\mu^2} exp\{-\dfrac{(a_{max} - \ddot{x}_l)^2}{2\mu^2}\} & \ddot{x}_l < a_{max} \\ 0 & \ddot{x}_l \geq a_{max} \end{cases} \tag{12}$$

where a_{max} is the maximum positive acceleration and $\mu > 0$ is a constant. The mean value and the variance of the random acceleration $\ddot{x}_l$ are

$$E[\ddot{x}_l] = a_{max} - \sqrt{\frac{\pi}{2}}\mu \quad \text{and } var[\ddot{x}_l] = \frac{4-\pi}{2}\mu^2, \tag{13}$$

respectively. Similarly, when $\ddot{x}_l < 0$, the probability density function has the form

$$P_r(\ddot{x}_l) = \begin{cases} \dfrac{(\ddot{x}_l - a_{-max})}{\mu^2} \exp\{-\dfrac{(\ddot{x}_l - a_{-max})^2}{2\mu^2}\} & \ddot{x}_l > a_{-max} \\ 0 & \ddot{x}_l \leq a_{-max} \end{cases} \tag{14}$$

where a_{-max} is negative. The mean value and the variance of the random acceleration $\ddot{x}_l$ are

$$E[\ddot{x}_l] = a_{-max} + \sqrt{\frac{\pi}{2}}\mu \quad \text{and } var[\ddot{x}_l] = \frac{4-\pi}{2}\mu^2, \tag{15}$$

respectively. The state equation in (9) can be rewritten as

$$x_l = F_{l|l-1}x_{l-1} + U_{l-1}\bar{a}_{l-1} + w_{l-1}. \tag{16}$$

where

$$F_{l|l-1} = \begin{bmatrix} 1 & T & \frac{1}{\alpha^2}\left(-1 + \alpha T + e^{-\alpha T}\right) \\ 0 & 1 & \frac{1}{\alpha}\left(1 - e^{-\alpha T}\right) \\ 0 & 0 & e^{-\alpha T} \end{bmatrix} \tag{17}$$

$$U_{l-1} = \begin{bmatrix} \frac{1}{\alpha}\left(-T + \frac{\alpha T^2}{2} + \frac{1 - e^{-\alpha T}}{\alpha}\right) \\ T - \frac{1 - e^{-\alpha T}}{\alpha} \\ 1 - e^{-\alpha T} \end{bmatrix} \tag{18}$$

By choosing T to be small enough such that αT is small, $F_{l|l-1}$ can be reduced to

$$F_{l|l-1} = \begin{bmatrix} 1 & T & T^2/2 \\ 0 & 1 & T \\ 0 & 0 & 1 \end{bmatrix} \triangleq F, \tag{19}$$

by employing the Taylor series of e^x and ignoring higher-order small terms. It is seen from (19) that $F_{l|l-1}$ is independent of l and the choice of α. Similarly, U_l can be approximated as $U_{l-1} = [0, 0, \alpha T]^T$.

It is clear that the CS model involves three parameters α, a_{max}, and a_{-max}. These parameters need to be predetermined in maneuvering target tracking. If some large values are chosen for these three parameters, the CS model performs well in tracking strong maneuvering targets which have large accelerations and vastly time-varying motion models, but has low accuracy in tracking nonmaneuvering targets which have small accelerations and insignificantly time-varying motion models. On the contrary, small values of α, a_{max}, and a_{-max} may improve the accuracy of the nonmaneuvering target tracking, but lead to poor performance in tracking strong maneuvering targets. In light of this, we introduce a scaling factor which can be adaptively adjusted to improve the tracking performance for both maneuvering and nonmaneuvering targets.

The norm of the residual is defined by $D_l = v_l^T P_l^{-1} v_l$, where $v_l = \left(z_l - \hat{z}_{l|l-1}\right)$ denotes the residual, and P_l denotes the covariance matrix of the residual at the time instant l.

For maneuvering targets, the difference between D_l and D_{l-1} is large. Otherwise, the difference is small. We define a scaling factor as

$$\lambda_l = \ln\left(D_l / D_{l-1} + 1\right), \tag{20}$$

and therefore, the previous parameters at the instant of l is adjusted as:

$$\alpha_l = \lambda_l \alpha_0,\ a_{max,l} = \lambda_l a_{max,0},\ \ a_{-max,l} = \lambda_l a_{-max,0} \tag{21}$$

where α_0, $a_{max,0}$ and $a_{-max,0}$ are initial parameters.

For the case where the acceleration of the target is large, the value of λ_l is also large, and therefore, α_l, $a_{max,l}$ and $a_{-max,l}$ are large enough to adapt to the target maneuvering.

Otherwise, a small value λ_l is chosen when the acceleration of the target is small with the goal of improving the accuracy of target tracking. This improved CS model is referred to as the adaptive CS (ACS) model in the following part.

B. CUBATURE KALMAN FILTER

For this nonlinear measurement model, a suboptimal solution for estimating the state vector can be achieved by employing the Cubature Kalman filter algorithm [26], [27]. As a result of (19), for small αT, the state vector can be updated by

$$\hat{x}_{l|l-1} = F\hat{x}_{l-1|l-1} + U_{l-1}\bar{a}_{l-1} \tag{22}$$

and its associated covariance is calculated as

$$P_{l|l-1} = FP_{l-1|l-1}F^T + Q_{l-1}. \tag{23}$$

On the measurement update stage, the cubature points ξ_i, $i = 1, \ldots, L$ are transformed so that they can capture the mean and covariance of the predicted density $p(x_l|z_{l-1}) = \mathcal{N}\left(x_l; \hat{x}_{l|l-1}, P_{l|l-1}\right)$. The transformed points are

$$\boldsymbol{\delta}_{i,l|l-1} = \sqrt{\boldsymbol{P}_{l|l-1}}\xi_i + \hat{x}_{l|l-1}, \quad i = 1, \ldots, L \tag{24}$$

where $\sqrt{P_{l|l-1}}$ is the matrix square root of $P_{l|l-1}$, that is, $\left(\sqrt{P_{l|l-1}}\right)\left(\sqrt{P_{l|l-1}}\right)^T = P_{l|l-1}$. The set of cubature points ξ_i is defined by $\xi_i = \sqrt{n_x}[I_{n_x} \vdots -I_{n_x}]_i$, where $\vdots$ denotes matrix concatenation, $[A]_i$ denotes the i-th column of matrix A, and I_{n_x} is an identity matrix of size n_x. The total number of points required in a CKF is $L = 2n_x$. Hence, a highly desirable property of the CKF is that the required number of cubature points increases only linearly with the dimension of the state vector.

These transformed cubature points are then propagated through the measurement function $\boldsymbol{\beta}_{i,l|l-1} = h(\boldsymbol{\delta}_{i,l|l-1})$ which are used to compute the predicted bearing measurement as a weighted sum $\hat{z}_{l|l-1} = \sum_{i=1}^{L} w_i\boldsymbol{\beta}_{i,l|l-1}$, where $w_i = \frac{1}{2n_x}, i = 1, 2, \ldots, L$ are the cubature weights. The Kalman gain can be obtained as $G_l = P_{xz,l|l-1}P_{zz,l|l-1}^{-1}$ where

$$P_{zz,l|l-1} = \sum_{i=1}^{L} w_i\left(\boldsymbol{\beta}_{i,l|l-1} - \hat{z}_{l|l-1}\right)\left(\boldsymbol{\beta}_{i,l|l-1} - \hat{z}_{l|l-1}\right)^T + R_l \tag{25}$$

is the innovation covariance and

$$P_{xz,l|l-1} = \sum_{i=1}^{L} w_i\left(\boldsymbol{\delta}_{i,l|l-1} - \hat{x}_{l|l-1}\right)\left(\boldsymbol{\beta}_{i,l|l-1} - \hat{z}_{l|l-1}\right)^T \tag{26}$$

is the cross covariance matrix.

With a new measurement z_l entering the algorithm, the predicted state is updated using the Kalman gain as

$$\hat{x}_{l|l} = \hat{x}_{l|l-1} + G_l\left(z_l - \hat{z}_{l|l-1}\right) \tag{27}$$

and the corresponding error covariance is given by

$$P_{l|l} = P_{l|l-1} - G_lP_{zz,l|l-1}G_l^T. \tag{28}$$

For simplicity, the Cubature Kalman filter based on the ACS model is referred to as adaptive Cubature Kalman filter in the paper.

IV. NUMERICAL SIMULATIONS AND PERFORMANCES ANALYSIS

In this section, the performance of the DCBF for wideband signals is compared with that of the CBF and the MVDR. In addition, the adaptive CS model for tracking a maneuvering target is also numerically evaluated.

A. PERFORMANCE EVALUATION OF DCBF

We consider a circular array with radius $r = 0.5$ m consisting of 80 microphones evenly distributed on the circumference. Assume that SNR is 10 dB and the sound speed $c = 340$ m/s. The azimuth angle and the elevation angle are both 0°. For the sake of comparison, the DOA estimates produced by the CBF and the MVDR are also given, respectively. The beam power is respectively calculated at a given direction for each kr. The number of iterations of the R-L algorithm is 20.

Fig. 2(a) shows the CBF beam output for kr ranging from 3 to 15. As expected, the CBF has a wide beamwidth at low kr. Fig. 2(b) and Fig. 2(c) respectively depict the outputs of the MVDR and the DCBF. It is seen from Fig. 2 that the DCBF yields a narrower beamwidth and lower sidelobe levels when compared to the CBF. In the wideband DOA estimation, we consider a frequency range from 300 to 3400 Hz which is contained in the Human voice frequency range. The sampling frequency is 8 kHz. And a 1024-point fast Fourier transform (FFT) is used for the spectral analysis of the time-series. The RL algorithm runs with 20 iterations for each frequency bin. The MVDR employs 160 snapshots while the CBF and the DCBF only employ 10 snapshots. The azimuth angle is 0°. Fig. 3(a), (b) and (c) show the outputs of the CBF, MVDR, DCBF for different frequencies. It is seen from Fig. 3 that both the DCBF and the MVDR have narrow beamwidths. Fig. 4 shows the average beampattern of the three beamforming techniques. As illustrated in Fig. 4, the DCBF performs the best in terms of beamwidth and sidelobe levels, even though the DCBF uses fewer snapshots than the MVDR. It is worth mentioning that the outputs of the DCBF for all frequency bins are uniformly averaged in the wideband DCBF algorithm.

In order to illustrate the superior directivity of the DCBF, we consider the case where one signal source and one interference source are close in terms of directions. Specifically, the azimuth angles of signal and interference are 0° and 3°, respectively. As shown in Fig. 5, the CBF and the MVDR cannot identify two sources, but the DCBF can effectively distinguish them.

B. ADAPTIVE CS MANEUVERING MODEL TESTING

In this section, the adaptive CS model is compared with the conventional CS model in the context of tracking a maneuvering target by using the CKF. As shown in Fig. 6(a), three receivers are located at (0 m, 0 m), (4000 m, 6000 m)

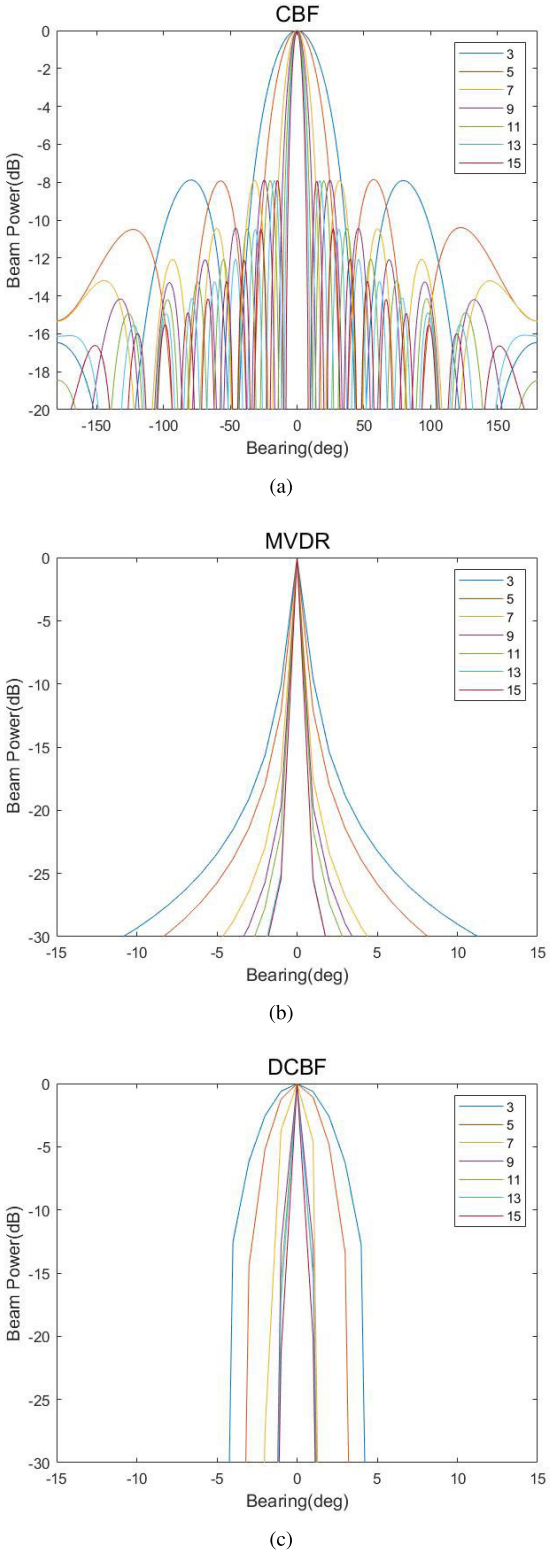


FIGURE 2. Beamforming outputs with different kr. (a) CBF; (b) MVDR; (c) DCBF.

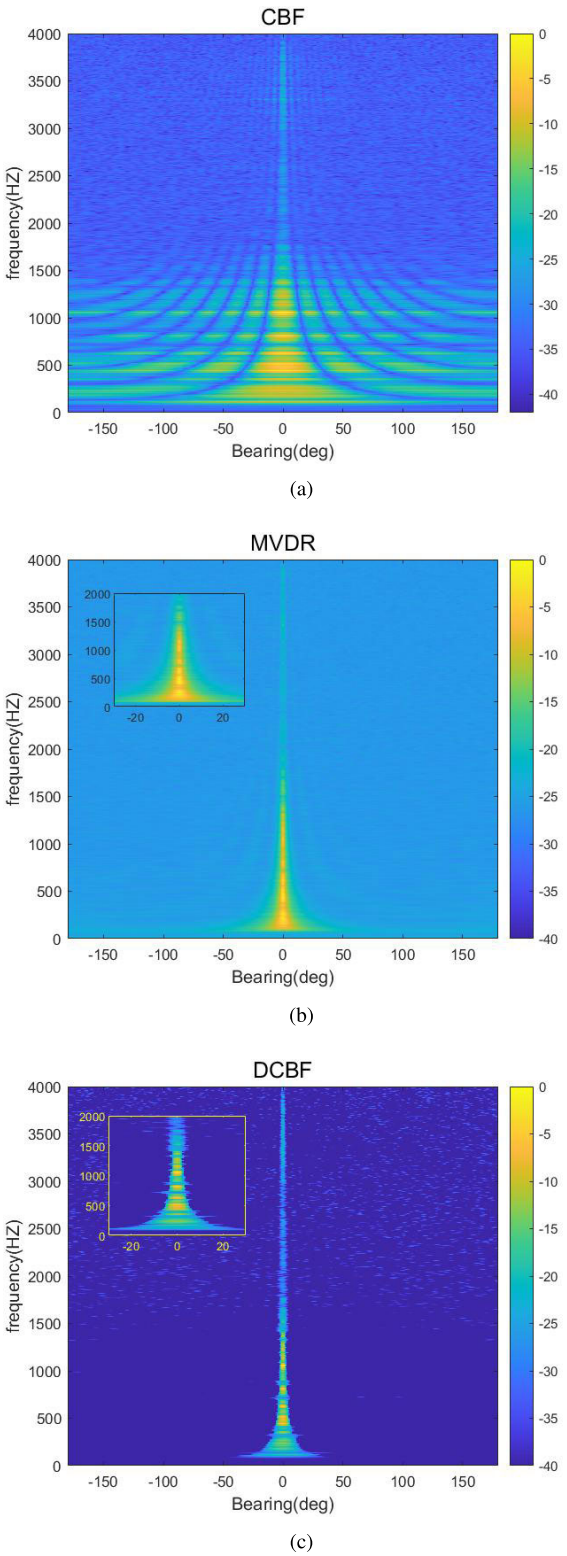


FIGURE 3. Beam power outputs of all frequency bins. (a) CBF; (b) MVDR; (c) DCBF.

and (7000 m, 5000 m), respectively. The standard deviation of measurement of bearing is 5°. The initial position of the target is located at (3000 m, 7000 m). The sampling time interval is 1 s and the total tracking time is 1200 s, the motion of the target can be divided into five stages. During the period of 0∼300 s, it moves with a unform velocity

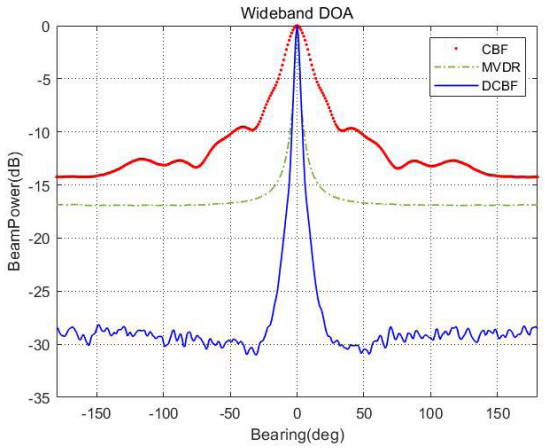


FIGURE 4. The wideband DOA estimations by the CBF, the MVDR and the DCBF.

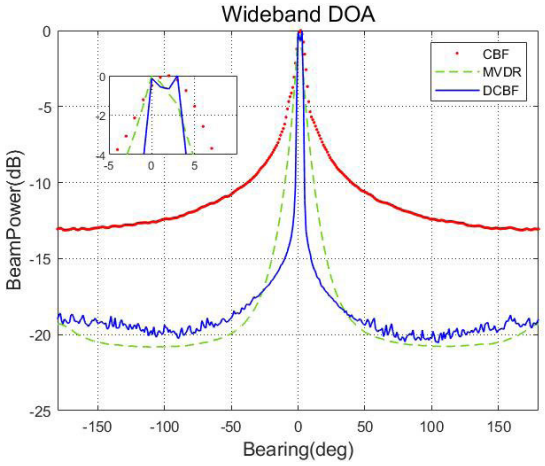


FIGURE 5. Resolution comparison among the CBF, the MVDR and the DCBF.

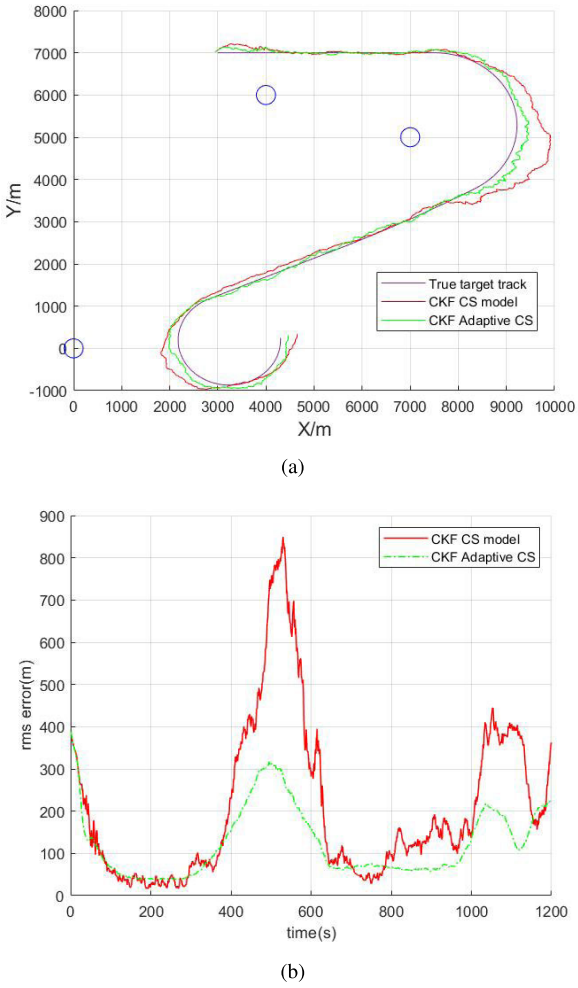


FIGURE 6. Maneuvering target tracking with the CS model and the ACS model. (a) True and estimated trajectories; (b) RMSE in position. The circles denote the positions of three receivers.

($x = 15$ m/s, $y = 0$ m/s), which is the stage I. Then it turns right at a rate of 0.5°/s for 300 s, which is the stage II. From 601∼800 s, it moves with a uniform acceleration of ($\ddot{x} = 0.02$ m/s^2, $\ddot{y} = 0$ m/s^2), which is the stage III. After that, it moves with a uniform velocity during 801∼950 s, which is the stage IV. During the final period of 250 s, it turns left at a rate of 0.5°/s, which is the stage V. The parameters $\alpha_0 = 1/60$, $a_{max,0} = 0.05$ m/s^2 and $a_{-max,0} = -0.05$ m/s^2.

Fig. 6(a) depicts the true and estimated trajectories. Fig. 6(b) shows the root-mean square errors (RMSEs) of the CKF with the CS model and the ACS model. The number of Monte Carlo runs is 500. It is seen from Fig. 6(a) that two models can both track the target in the five stages. But, when the target turns right or left, the tracking trajectories of two models deviate the true trajectory as shown in Fig. 6(a) to different extents. From Fig. 6(b), we can see that the ACS model has a smaller RMSE than the CS model. In particular, the tracking RMSE of the ACS model can be as small as half of that of the CS model at time equal to 500 s. Thus, the ACS model outperforms the CS model in tracking maneuvering targets.

Further, the robustness of the ACS model is evaluated using different initial parameters. Table 1 presents the overall average tracking RMSEs of two models with different α_0 and a_{max}. In each table cell, the values before and after the slash correspond the CS model and the ACS model, respectively. The notation Inf in the table means that the RMSE is too large. From Table 1, we can see that the ACS model is better than the CS model since the former has small RMSEs. Meanwhile, when α_0 decreases, the RMSEs of the CKF with the two motion models both decrease due to the smaller step size in iteration. The ACS model has smaller RMSEs than the CS model due to the significant role of the scaling factor in tracking the strong maneuvering target. What's more, when α_0 and a_{max} deviate a lot from the true value, such as $\alpha_0 = 1/20$ $a_{max} = 0.07$, the target may be lost when using the CS model but the ACS model can track the target stablely.

Table 2 shows the tracking RMSEs for the target with different maneuvering acceleration. Here, stage III denotes

TABLE 1. Comparison between the ACS and the CS with different α_0 and a_{max}.

RMSE (CS/ACS) a_{max} / α_0	0.01	0.03	0.05	0.07
1/20	577.5/370.7	1086.1/279.1	1914.6/165.9	Inf/133.8
1/40	639.8/349.3	725/174.6	214.6/134.4	201.9/116.3
1/60	706.6/304	288.2/163	235.8/125.7	136.7/115.5
1/80	479.8/286.4	242.9/158.6	182.3/124.2	172.5/115.6
1/100	428.6/273.5	228.5/155.6	141.3/123.7	117.3/112.7

TABLE 2. Comparison between the ACS and the CS under different accelerations.

RMSE [1] stage / $\ddot{x}$	I	II	III	IV	V
0.1	67.8/68.8	180.1/127.0	59.4/63.7	72.6/74.3	1102.8/686.5
0.2	57.7/67.0	181.6/125.8	92.3/64.9	203.6/82.2	2467.5/1889.1
0.4	62.2/68.4	198.5/124.4	126.9/72.3	326.8/308.5	3793.5/2934.6

[1] CS/ACS

TABLE 3. Summary of processing algorithms.

DCBF	Narrower beamwidth when compared with the CBF; low sidelobe levels; robust to array position error; require the angle shift-invariance of beam pattern
CKF	Lower computational complexity than that of the PF; RMSE is smaller than the UKF; more stable than the UKF
Adaptive CS	Can stably track maneuvering or nonmaneuvering targets; the averaged RMSE is lower than CS

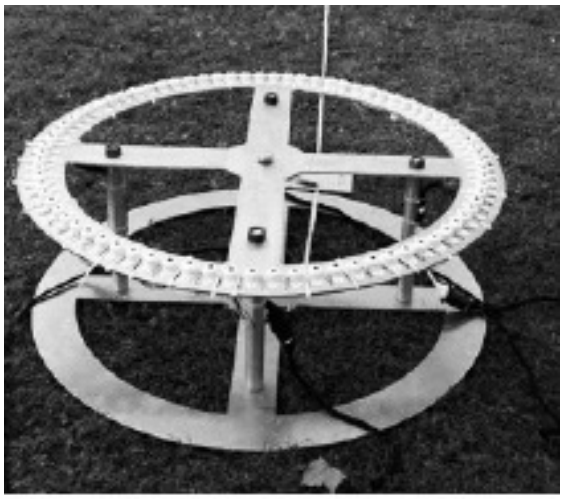
the third stage with a uniform acceleration. It is obvious that the ACS model outperforms the CS model when the target's acceleration becomes larger and larger. In addition, the tracking RMSEs for other stages are also provided in Table 2. Similarly, the ACS model performs well when the target turns right or left. In the first stage, when the target moves with a constant velocity, the CS model shows slightly better performance than that of the ACS model. But in the fourth stage where the estimated angle produced by the DCBF is not very accurate, the ACS model has smaller RMSEs than the CS model. This implies that the ACS model is more robust when the estimated angle is not very accurate. Table 3 summarizes some features of the DCBF, the CKF, and the adaptive CS model. Detailed discussion on the pros and cons of the classic tracking algorithms, such as KF, EKF, UKF, PF, CKF, IMM, AMM, can be found in [14], [28]–[30] and references therein.

V. EXPERIMENTAL RESULTS

Besides the numerical results, we also build a real distant speech perception system based on the proposed algorithm, and we test the proposed algorithm based on the experimental data. The experimental results show that when compared with some conventional algorithm, the proposed algorithm can achieve the same tracking performance with much lower computational complexity.

The distant speech perception system is shown in Fig. 7. The system consists of two microphone arrays, a signal processing hardware platform and an upper computer. The microphone rectangular array consists of eight horizontal arrays with ten equally spaced elements. The element spacing between two neighboring microphones is 3.8 cm. In the vertical direction, the space between two neighboring horizontal linear arrays is 3.8 cm. The circular array with radius 0.5 m consist of 80 elements with a uniform spacing of 0.04 m. As such, the beamforming algorithms implemented on the two arrays can be integrated into the hardware platform to capture distant speech signals in a real time manner, and the recorded data are processed by our proposed algorithms on the upper computer.

The omnidirectional microphone COTT-C5 has the sensibility of -45 dB and the frequency response from 20 Hz to 20 kHz. We developed an integrated pre-amplifier circuit



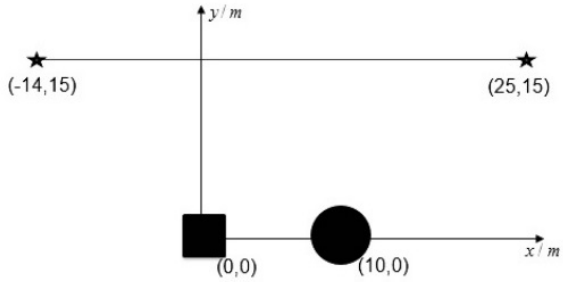
(a)

(b)

(c)

FIGURE 7. The hardware platform.(a) Circular array; (b) Rectangular array; (c) Processor cabinet and upper computer.

module to enhance the capability of microphones for sensing weak speech signals. The signal processing platform consists of a signal conditioning module, an A/D module, a main controller unit and a DC power supply module. The signal conditioning module consists of a amplifier with 20 dB gain and a bandpass filter with a passband from 285 Hz to 3.7 kHz. The signal conditioning module connects the A/D module with a differential signal transmission cable. In the A/D module,



(c)

FIGURE 8. The experiment configuration.

the chips ADS1601 are used whose sampling rate is 48 kHz. The Z-turn board based main controller unit consists of two ARM Cortex-A9 processing systems and a Programmable Logic unit, which are utilized for 80-channel synchronous sampling, external communication, and the implementation of beamforming etc. Meanwhile, the 80-channel data are transmitted to the upper computer through RS-232 serial ports, where the proposed wideband DCBF algorithm and the improved tracking algorithms are implemented to the received data for distant speech reception.

It is worth mentioning that the rectangular array is only utilized to assist the circular array to estimate the initial position of the speech of interest. Specifically, the CBF is applied to the rectangular array to estimate the DOA of the speech of interest. Meanwhile, the DOA of the speech of interest is also estimated by the DCBF implemented on the circular array. With the known positions of these two arrays and their estimates of the DOA, we can estimate the initial position of the speech source based on the Least Square method. With the estimated initial position of the speech source, we are able to track the speech source in a real-time manner by solely using the circular array with our proposed algorithms, which will be illustrated in Fig. 9.

The system performance is tested in outdoor experiments. The Fig. 8 shows the experimental setup. The centers of the rectangular and circular array are respectively located at (0 m, 0 m) and (10 m, 0 m) in the Cartesian coordinate system. The speech source moves along a straight line from (−14 m, 15 m) to (25 m, 15 m). The speech of interest is a record of an English lecture. The outdoor environmental noise level is in the range from 53 to 55 dB. A 512-point DFT is used in the wideband speech signal DOA estimation. In addition, $\alpha_0 = 1/60$ and $a_{max,0} = 0.1$ m/s^2 are chosen for the ACS model.

The cubature Kalman filter is combined with the ACS model (ACKF) to track a maneuvering speech source. The speech source moves along a straight line from (25 m, 15 m) to (−5 m, 15 m) and stops at the position (10 m, 15 m) for a while. For the sake of comparison, the results of the PF with the ACS model (APF) are also provided. In the tracking algorithm, the initial positions are provided by the weighted least square (LS) method with the estimated DOAs. It is seen from

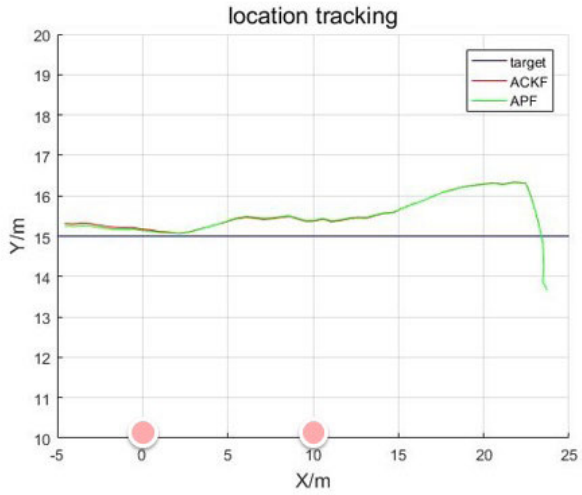


FIGURE 9. Position tracking of a maneuvering speech source using ACKF and APF.

Fig. 9 that the ACKF and the APF can both stably track the position of the speech source. And two tracking trajectories almost overlap. The number of particles in the APF is 1000, while the ACKF only uses 8 cubature points. Thus, the CKF can achieve almost the same tracking performance as that of the PF but with a much lower computational complexity.

VI. CONCLUSION

We have derived the wideband DCBF algorithm and designed an adaptive cubature Kalman filter for a distant speech perception system. The DCBF algorithm is extended from narrow band to wideband signals which is employed to estimate of the DOA of the speech signal of interest using a circular microphone array. With the adaptive CS motion model, it has been shown that the cubature Kalman filter can reliably track both maneuvering and nonmaneuvering speech source with low computational complexity. By integrating the speech enhancement approach and the improved tracking algorithm into the hardware processing platform, we can capture the distant speech and track the speech source movement in a real-time manner, which has been demonstrated by the outdoor experiments.

ACKNOWLEDGMENT

The authors thank the helpful contributions from Prof. TC. Yang.

REFERENCES

[1] K. Kumatani, J. McDonough, and B. Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 127–140, Nov. 2012.

[2] P. Swietojanski, A. Ghoshal, and S. Renals, "Convolutional neural networks for distant speech recognition," *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 1120–1124, Sep. 2014.

[3] K. Niwa, Y. Hioka, and K. Kobayashi, "Optimal microphone array observation for clear recording of distant sound sources," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 10, pp. 1785–1795, Oct. 2016.

[4] S. A. Khoubrouy and J. H. L. Hansen, "Microphone array processing strategies for distant-based automatic speech recognition," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1344–1348, Oct. 2016.

[5] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. London, U.K.: Springer-Verlag, 2010.

[6] E. Nemer and W. Leblanc, "Single-microphone wind noise reduction by adaptive postfiltering," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2009, pp. 177–180.

[7] E. A. P. Habets, J. Benesty, and P. A. Naylor, "A speech distortion and interference rejection constraint beamformer," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 3, pp. 854–867, Mar. 2012.

[8] B. Schwartz, S. Gannot, and E. A. P. Habets, "Online speech dereverberation using Kalman filter and EM algorithm," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 2, pp. 394–406, Feb. 2015.

[9] F. Daum, "Nonlinear filters: Beyond the Kalman filter," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 20, no. 8, pp. 57–69, Aug. 2005.

[10] S. Sadhu, S. Mondal, M. Srinivasan, and T. K. Ghoshal, "Sigma point Kalman filter for bearing only tracking," *Signal Process.*, vol. 86, no. 12, pp. 3769–3777, Dec. 2006.

[11] K. Wu, V. G. Reju, A. W. H. Khong, and S. T. Goh, "Swarm intelligence based particle filter for alternating talker localization and tracking using microphone arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 6, pp. 1384–1397, Jun. 2017.

[12] S. Bordonaro, P. Willett, Y. Bar-Shalom, and T. Luginbuhl, "A Gaussian-sum based cubature Kalman filter for bearings-only tracking," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 49, no. 2, pp. 1161–1176, Apr. 2013.

[13] H. Zhou and K. S. P. Kumar, "A 'current' statistical model and adaptive algorithm for estimating maneuvering targets," *J. Guid., Control, Dyn.*, vol. 7, no. 5, pp. 596–602, Sep. 1984.

[14] X. Rong Li and V. P. Jilkov, "Survey of maneuvering target tracking. Part I. Dynamic models," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 39, no. 4, pp. 1333–1364, Oct. 2003.

[15] Z. Zhuoran, Y. Guangqiang, and Z. Xiaolin, "Adaptive algorithm based on modified current statistical model for passive tracking," *Comput. Eng. Appl.*, vol. 53, no. 3, pp. 124–130, 2017.

[16] L. Hui, S. Ying, Z. An, and C. Cheng, "A new adaptive filtering algorithm in maneuvering target tracking," *J. Northwestern Polytech. Univ.*, vol. 24, no. 3, pp. 354–357, 2006.

[17] R. Visina, Y. Bar-Shalom, and P. Willett, "Multiple-model estimators for tracking sharply maneuvering ground targets," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 3, pp. 1404–1414, Jun. 2018.

[18] M. Eltoukhy, M. O. Ahmad, and M. N. S. Swamy, "An adaptive turn rate estimation for tracking a maneuvering target," *IEEE Access*, vol. 8, pp. 94176–94189, 2020.

[19] L. Xu, X. Rong Li, and Z. Duan, "Hybrid grid multiple-model estimation with application to maneuvering target tracking," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 52, no. 1, pp. 122–135, 2016.

[20] K. Kumatani, T. Arakawa, K. Yamamoto, J. McDonough, R. Raj, B. Singh, and I. Tashev, "Microphone array processing for distant speech recognition: Towards real-world deployment," in *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2012, pp. 1–10.

[21] C. Zhang, T. Yu, and J. H. L. Hansen, "Microphone array processing for distance speech capture: A probe study on whisper speech detection," in *Proc. Conf. Rec. 4th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2010, pp. 1707–1710.

[22] T. Yang, "Deconvolved conventional beamforming for a horizontal line array," *IEEE J. Ocean. Eng*, vol. 43, no. 1, pp. 160–172, Jan. 2018.

[23] M. Wax, T.-J. Shan, and T. Kailath, "Spatio-temporal spectral analysis by eigenstructure methods," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 4, pp. 817–827, Aug. 1984.

[24] J. A. O'Sullivan, R. E. Blahut, and D. L. Snyder, "Information-theoretic image formation," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2094–2123, Oct. 1998.

[25] R. E. Blahut, *Theory of Remote Image Formation*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[26] I. Arasaratnam and S. Haykin, "Cubature Kalman filters," *IEEE Trans. Autom. Control*, vol. 54, no. 6, pp. 1254–1269, Jun. 2009.

[27] P. Leong, S. Arulampalam, T. Lamahewa, and T. Abhayapala, "A Gaussian-sum based cubature Kalman filter for bearings-only tracking," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 49, no. 2, pp. 1161–1176, Apr. 2013.

[28] X. R. Li and V. P. Jilkov, "Survey of maneuvering target tracking. Part V. Multiple-model methods," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 41, no. 4, pp. 1255–1321, 2005.
[29] S. Chen, "Kalman filter for robot vision: A survey," *IEEE Trans. Ind. Electron.*, vol. 59, no. 11, pp. 4409–4420, 2011.
[30] A. Akca and M. Ö. Efe, "Multiple model Kalman and particle filters and applications: A survey," *IFAC-PapersOnLine*, vol. 52, no. 3, pp. 73–78, 2019.

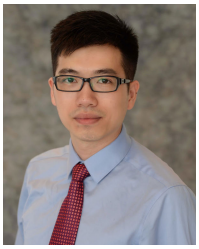
XIANG PAN received the bachelor's degree in underwater acoustic electronic engineering from the Harbin Ship Engineering Institute, in 1989, the master's degree in underwater acoustic engineering from the China Ship Research and Development Academy, in 1998, and the Ph.D. degree in information and communication engineering from Zhejiang University, in 2003.

He was a Visiting Scholar with Concordia University, Canada, from June 2009 to September 2009, the University of Victoria, Canada, from April 2011 to April 2012, and the University of Connecticut, USA, from May 2014 to September 2014. He is currently an Associate Professor with Zhejiang University. His research interests include statistical signal processing, acoustic signal processing, pattern recognition, and image processing.

YUE BAO received the B.Eng. degree in information engineering and the M.Eng. degree in information and electronic engineering from Zhejiang University, Hangzhou, China, in 2015 and 2018, respectively.

YITING ZHU received the B.Eng. degree in information engineering and the M.Eng. degree in information and communication engineering from Zhejiang University, Hangzhou, China, in 2015 and 2018, respectively.

HUANGYU DAI received the B.Eng. degree in electronic information science and technology from Dalian Maritime University, Dalian, China, in 2019. He is currently pursuing the D.Eng. degree in information and communication engineering with Zhejiang University, Hangzhou, China.



JIANGFAN ZHANG (Member, IEEE) received the B.Eng. degree in communication engineering from Huazhong University of Science and Technology, Wuhan, China, in 2008, the M.Eng. degree in information and communication engineering from Zhejiang University, Hangzhou, China, in 2011, and the Ph.D. degree in electrical engineering from Lehigh University, Bethlehem, PA, USA, in 2016.

From 2016 to 2018, he was a Postdoctoral Research Scientist with the Department of Electrical Engineering, Columbia University, New York, NY, USA. Since 2018, he has been with the Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, MO, USA, where he is currently an Assistant Professor. His research interests include signal processing, machine learning, and their applications to cybersecurity, the Internet of Things, sensor networks, smart grid, and sonar processing. He was a recipient of the Dean's Doctoral Student Assistantship, the Gotshall Fellowship, and the P. C. Rossin Doctoral Fellowship at Lehigh University.