Automatic Beamforming for Blind Extraction of Speech From Music Environment Using Variance of Spectral Flux-Inspired Criterion

Tao Yu, *Student Member, IEEE*, and John H. L. Hansen, *Fellow, IEEE*

***Abstract*—This paper addresses the problem of automatic beamforming for blind extraction of speech in a music environment, using multiple microphones. A new criterion is proposed based on the variance of the spectral flux (VSF), which is shown to be a compound measure of the Kurtosis and across-time correlation for the time–frequency domain signals. Spectral flux (SF) had been adopted as a feature that distinguishes speech from other acoustic noises and the VSF of speech tends to be larger than that of other acoustic sounds. Henceforth, maximization of VSF can be employed as one potential criterion to identify the speech direction-of-arrival (DOA), in order to extract speech from the noisy observations. We construct a VSF-inspired cost function and develop a complex-value fixed-point algorithm for the optimization. Then, the stability of the proposed algorithm is analyzed based on the second-order Taylor series expansion. Rather than the DOA identification ambiguity caused by subspace decomposition-based methods or maximization of non-Gaussianity-based approaches, both real and simulated evaluations indicate that the VSF-inspired criterion can effectively extract speech from a music diffuse noise field or a musical interference noise field. A key feature of the proposed approach is that it can operate blindly, i.e., it does not require *a priori* knowledge about the array geometry, the noise covariance matrix, or the geometrical knowledge of the location of desired speech. Therefore, this study offers a potential perspective for blindly extracting speech from a music environment.**

***Index Terms*—Array signal processing, blind beamforming, blind source extraction (BSE), speech enhancement.**

I. Introduction

MICROPHONE array beamforming has been widely and extensively studied for teleconferencing, speech enhancement, speech recognition, and hearing aids [1]. Aiming at removing unwanted interference and noise from a desired speech, beamforming techniques exploit spatial and spectral diversity to discriminate between desired and undesired signal components. However, beamforming has not had the success we hope for; acoustic beamforming typically assumes that the array steering vector of the desired speech as well as the interference-plus-noise covariance matrix are known beforehand, which is generally impractical for real-world applications.

Due to the uncertainty of *a priori* knowledge of the array steering vector, a great diversity in blind array processing algorithms that use an *a posteriori* style data-driven approaches can be found in literature, and are shown to be more robust in practical environments. This kind of approach is called *blind beamforming*, which automatically adjusts the beampattern to achieve the best reception of the signal of interest (SOI) without explicit knowledge of the array shape, the direction of arrival (DOA) of SOI, or interference-plus-noise covariance matrix. Generally, we can divide these blind beamforming approaches into two categories: second-order statistics (SOS)-based and higher order statistics (HOS)-based approaches.

The SOS-based blind beamformer explores the eigenspace between the interference-plus-noise covariance matrix and the array observation data covariance matrix (or the desired speech covariance matrix) [2]–[4]. Under the assumption that the desired speech has a dominant power over the interference or noise in all the frequencies, the principle eigenvector obtained through a generalized eigenvalue decomposition (GSVD) of the interference-plus-noise covariance matrix and the array observation data covariance matrix is shown to be equivalent to the weights of the maximal signal-to-noise power ratio (Max-SNR) beamformer, and only differs by a scalar factor from the minimal variance distortionless response (MVDR) beamformer [4]. Although this approach is simple, it requires perfect knowledge of the interference-plus-noise covariance matrix, which is generally unavailable in practice. Moreover, considering the sparsity of speech in the time–frequency domain, the assumption of dominant power of the desired speech over noise is not always guaranteed.

Another set of blind acoustic beamforming solutions considers the super-Gaussianity of the speech probability distribution in the time–frequency domain, and therefore the HOS are used. For example, the Kurtosis approaches used in [5] and [6] automatically identify the array steering vector by finding the DOA that has the local maximal Kurtosis. While these approaches can successfully identify the DOA of the desired speech, they require that all the noises to be Gaussian and the array structure must be known. In a non-Gaussian noise field, multiple local maxima of Kurtosis would cause ambiguous identification of the desired speech DOA.

Blind source separation (BSS), blind source extraction (BSE), or independent component analysis (ICA) are other

Manuscript received November 05, 2009; revised January 29, 2010; accepted March 05, 2010. Date of publication August 26, 2010; date of current version September 15, 2010. This work was supported in part by the AFRL through a subcontract to RADC, Inc., under Grant FA8750-09-C-0067 and in part by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. Hansen. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Maurizio Omologo.

The authors are with the Center for Robust Speech System (CRSS), University of Texas at Dallas, Richardson, TX 75083-0688 USA (e-mail: tao.yu@student.utdallas.edu; john.hansen@utdallas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2010.2069790

approaches that stand out in order to separate unknown sources from the observed mixtures based solely on their HOS, without specifically knowing the array structure and DOAs of each source signals. However, the main limitation of blind source separation is the existence of ambiguities in the independence criterion, namely the scaling and permutation ambiguity problem. To overcome this problem, a set of geometrically constrained BSS (GSS) algorithms have been proposed [7] and further explored and incorporated in subsequent studies [8]–[11]. By assuming that the desired speech comes from a certain range of DOAs, the desired speech can be extracted through selecting or constraining the ICA separation weights pointing to a preselected DOA range without causing permutation ambiguity, and as an additional benefit, a faster convergence rate can be achieved due to the constrained searching space. However, this semi-blind solution still suffers from the requirement of *a priori* DOA knowledge of the desired speech.

In this paper, we consider an acoustic blind beamforming application with only one desired speaker exists but with unknown directional music interferences simultaneously activating in an noisy environment. We construct a time–frequency domain beamformer that can blindly extract the desired speech. In contrast to conventional approaches, we identify the DOA of the desired speech via maximal variance of spectral flux criterion (VSF) in the signal subspace, without assuming any *a priori* information. Spectral flux (SF) and the VSF have been successfully adopted as effective features that distinguish speech from music [12], [13]. Assuming that only one speech source is active in a musical interferences or diffuse noisy environment, we show that thought the maximization of VSF at the beamformer's output, speech components can be effectively extracted.

This paper is organized as follows. In Section II, a microphone array-based signal model is presented in the time–frequency domain and the classical optimal beamforming theory is briefly reviewed. From a signal subspace perspective, the general blind beamforming procedures are described in Section III, and in Section IV we propose our algorithm followed by its stability analysis. Section V presents a batch processing implementation of the proposed algorithm for speech enhancement. We evaluate and compare our proposed method with other techniques in Section VI. Finally, in Section VII, we draw the conclusion and suggest directions for future research.

II. Background

A. Microphone Array Based Signal Model

Consider one desired speech impinges on a array of M microphones. Taking the short-time Fourier transform (STFT) of the time domain signal, the signal model in each time-frame and frequency-bin can be written as

$$\boldsymbol{x}(t,k) = \boldsymbol{a}(t,k)s(t,k) + \boldsymbol{i}(t,k) + \boldsymbol{n}(t,k) \tag{1}$$

where $\boldsymbol{x} \in C^{M\times 1}$ is the array observation data vector, $s \in C$ is the desired speech, $\boldsymbol{a} \in C^{M\times 1}$ is the unknown (maybe time-varying) array steering vector, $\boldsymbol{i} \in C^{M\times 1}$ represents a collection of interference signals, $\boldsymbol{n} \in C^{M\times 1}$ is the background noise vector, and t and k are the time-frame index and frequency-bin index, respectively. Commonly, we can process each frequency-bin independently; thus, the notation of the frequency-bin index k is be omitted for briefness.

Assuming that each vector components of the model in (1) are mutually uncorrelated, the autocorrelation matrix for the observed data vector can be expressed as

$$\begin{aligned} R_{xx} &= E\{\boldsymbol{x}(t)\boldsymbol{x}(t)^H\} \\ &= R_{ss} + R_{[i+n][i+n]}, \\ &= R_{ss} + R_{ii} + R_{nn}, \\ &= \sigma_s^2 \boldsymbol{a}\boldsymbol{a}^H + \sum_{j=1}^{J-1} \sigma_j^2 \boldsymbol{f}_j \boldsymbol{f}_j^H + R_{nn}. \end{aligned} \tag{2}$$

where R_s, R_{ii} and R_{nn} are the autocorrelation matrices for the desired speech, the interferences and background noise, respectively, and σ_s^2 is the power of the desired speech. σ_j^2 is the jth interference power ($j \leq J-1$, i.e., the total number of interferences, and $J \leq M$) and $\boldsymbol{f}_j(k)$ is its corresponding array steering vector.

B. Optimal Beamforming

For a single frequency-bin, the optimal beamformer is a linear processor (filter) consisting of a set of complex weights [14]. The output of the beamformer is an estimate of the desired signal and is given by

$$y(t) = \hat{s}(t) = \boldsymbol{w}^H \boldsymbol{x}(t). \tag{3}$$

The weights are chosen according to some optimization criterion, such as the minimum mean square error (MMSE), the minimum variance distortionless response (MVDR), or the maximum signal-to-noise ratio (Max-SNR). Generally, the optimal weights have the same structure [14], as

$$\boldsymbol{w}_o = \mu R_{[i+n][i+n]}^{-1} \boldsymbol{a} \tag{4}$$

where μ is a scale factor decided by the optimization criterion. In practical arrays, however, the optimum weights are hardly obtained for two reasons:

- uncertainty of the array steering vector $\boldsymbol{a}$;
- uncertainty of the interference-plus-noise autocorrelation matrix $R_{[i+n][i+n]}$.

III. General Blind Beamforming Procedure

A. Signal Subspace

Consider the eigenvalue decomposition (EVD) of the autocorrelation matrix of the array observation data, which can be expressed as

$$R_{xx} = \sum_{m=1}^{M} \lambda_m \boldsymbol{u}_m \boldsymbol{u}_m^H \tag{5}$$

with eigenvalues ordered as $\lambda_1 > \lambda_2 > \cdots > \lambda_M$ and $\boldsymbol{u}_m$ is the eigenvector associated with the mth eigenvalue λ_m. We can then rewrite (5) as

$$R_{xx} = U_s \Lambda_s U_s^H + U_n \Lambda_n U_n^H \tag{6}$$

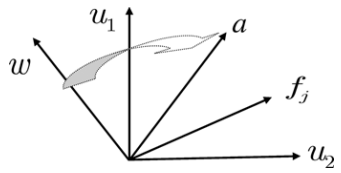

Fig. 1. Illustration of vectors in signal subspace.

where the matrix $U_s = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_J] \in C^{M\times J}$ contains the eigenvectors corresponding to the J ($J \leq M$) largest eigenvalues, and $\Lambda_s = diag\{\lambda_1, \ldots, \lambda_J\}$. The range space of U_s is called the *signal subspace* or the *signal-plus-interference subspace*. Its orthogonal complement is the *noise subspace* which is spanned by the columns of $U_n = [\boldsymbol{u}_{J+1}, \ldots, \boldsymbol{u}_M]$, and $\Lambda_n = diag\{\lambda_{J+1}, \ldots, \lambda_M\}$.

B. Blind Identification in Signal Subspace

To extract the desired speech, we need to select a weight vector $\boldsymbol{w}$ (normalized so that $\boldsymbol{w}^H\boldsymbol{w} = 1$) and generate the output as

$$y(t) = \boldsymbol{w}^H\boldsymbol{x}(t). \tag{7}$$

We would like to select $\boldsymbol{w}$ so that the interferences will be minimized and the desired signal component will be maximized. It is well known that U_s is spanned by $\boldsymbol{a}$ and $\boldsymbol{f}_j$; optimal $\boldsymbol{w}$ must lie in the signal subspace, as illustrated in Fig. 1. In some studies [3], [15] , it is assumed that the eigenvector corresponding to the maximal eigenvalues, namely the principal eigenvector, span the same space as $\boldsymbol{a}$ of the desired speech. Hence, in matched-filter design [15], the beamformer is given by

$$\boldsymbol{w}_{MF} = \boldsymbol{u}_1 \tag{8}$$

which has a beampattern with a maxima pointing in the direction of maximum coming power towards the array. Obviously, in the presence of interference, the direction of the maximum power does not necessarily equals to direction of the desired speech.

Therefore, we must search for the correct $\boldsymbol{w}$ by more sophisticated approaches. First, let us define the power normalized signal subspace as $\bar{U}_s = [\lambda_1^{-1/2}\boldsymbol{u}_1, \lambda_2^{-1/2}\boldsymbol{u}_2, \ldots \lambda_J^{-1/2}\boldsymbol{u}_J]$. Note that any vector that lies in $\bar{U}_s$ is given by

$$\boldsymbol{w} = \bar{U}_s\boldsymbol{r} \tag{9}$$

where $\boldsymbol{r} \in C^{J\times 1}$ and $\boldsymbol{r}^H\boldsymbol{r} = 1$ are unknown "rotation parameters." The output of the beamformer becomes

$$y(t) = \boldsymbol{w}^H\boldsymbol{x}(t) = \boldsymbol{r}^H\boldsymbol{z}(t) \tag{10}$$

where $\boldsymbol{z}(t) = \bar{U}_s^H\boldsymbol{x}(t)$, which transforms the array observation data vector $\boldsymbol{x}(t)$ into the signal subspace with normalized signal power, i.e.,

$$E\{\boldsymbol{z}(t)\boldsymbol{z}^H(t)\} = I_J. \tag{11}$$

and

$$E\{|y(t)|^2\} = E\{y(t)y^*(t)\} = 1. \tag{12}$$

We can search over all possible values of $\boldsymbol{r}$ until some measure specifically for the desired speech is maximized. Therefore, we must first define a cost function which measures the desired speech quality at the output response, as described in the following sections.

C. Maximal Non-Gaussianity Criterion

The principle of using non-Gaussianity to extract the desired speech is motivated by the central limit theorem, which loosely states that the sum of independent random variables with finite-support pdfs tends towards a Gaussian distribution, a classical basis stated in many ICA studies [16]. In real world applications, the mixture of various environmental sounds tend to be Gaussian; but speech has a super-Gaussian distribution. Therefore, one could choose $\boldsymbol{r}$ such that $y(t) = \boldsymbol{r}^H\boldsymbol{z}(t)$ has a non-Gaussian (or super-Gaussian) distribution.

Non-Gaussianity could be measured by HOS, such as the fourth order cumulant, i.e., the Kurtosis , which is vanished for a Gaussian distributed source. The Kurtosis of a zero-mean complex random variable y is defined as a real number [5], [17]

$$\mathrm{K}(y) \triangleq E\{|y(t)|^4\} - 2(E\{|y(t)|^2\})^2 - |E\{y^2(t)\}|^2 \tag{13}$$

and can be shown to be zero for any complex Gaussian variable. Notice that in the STFT domain, speech and many acoustic noises are generally circularly distributed [18], (i.e., independent and identically distributed (i.i.d.) of real and imaginary parts), hence $|E\{y^2(t)\}|^2 = 0$. Furthermore, if we impose the constraint that $E\{|y(t)|^2\} = 1$ as in (12), the Kurtosis can be calculated by

$$\mathrm{K}(y) = E\{|y(t)|^4\} - 2(E\{|y(t)|^2\})^2 = E\{|y(t)|^4\} - 2 \tag{14}$$

for circular signals, as used in [19].

Therefore, the cost function to solve the optimal weights can be constructed as

$$J_{\mathrm{NG}}(\boldsymbol{r}) = E\{|y(t)|^4\} = E\{|\boldsymbol{r}^H\boldsymbol{z}(t)|^4\} \tag{15}$$

and $\boldsymbol{r}$ can be solved by

$$\boldsymbol{r}_{\mathrm{NG}} = \arg\max_{\boldsymbol{r}^H\boldsymbol{r}=1} J_{\mathrm{NG}}(\boldsymbol{r}). \tag{16}$$

While maximization of non-Gaussianity can be applied to extract speech from the noise, it should be applied with caution. When the probability distribution of interferences or noise are not Gaussian, multiple extremum of the cost function in (16) can be found, resulting in DOA ambiguity. This problem has the same roots in the well known "permutation ambiguity" problem found in BSS [20], which aims at separating multiple sources.

IV. Maximal Variance of Spectral Flux Criterion

A. Spectral Flux and Its Variance

Spectral Flux (SF), a feature originally found to be useful in discriminating music from other acoustic sounds [12], [21], measures the ordinary Euclidean norm of the delta spectrum

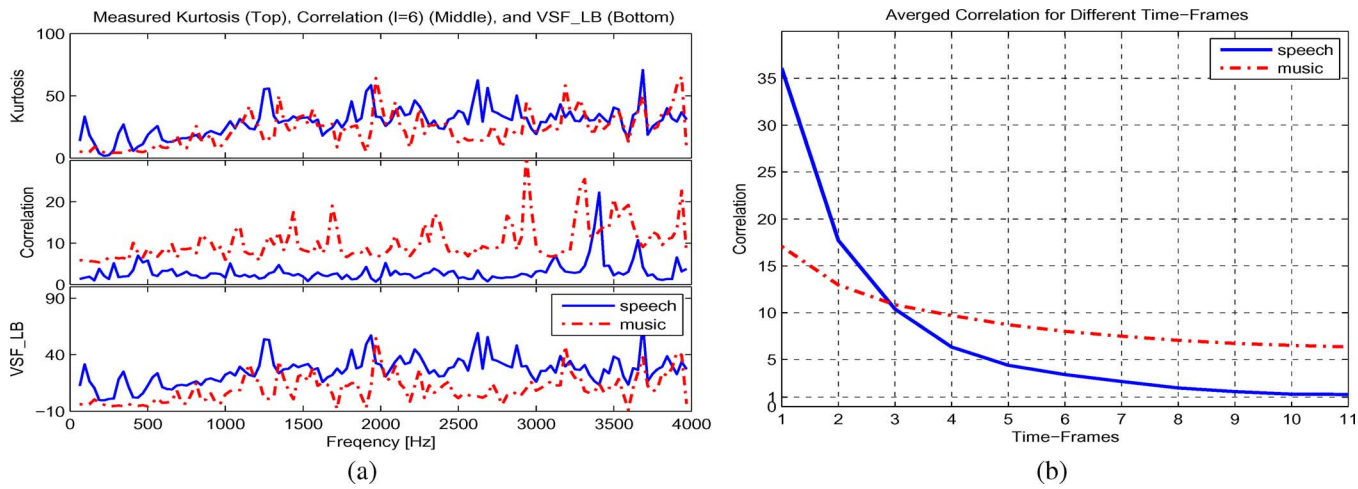

Fig. 2. (a) Illustration of relationship among the Kurtosis, correlation and VSF-LB. (b) Correlation drops as the time-frame difference l increases.

power (or magnitude) between two adjacent time-frames. For a practical implementation, the overlap of time-window used in the STFT transformation introduces a dependence between the spectra in consecutive frames; therefore, a general SF measure is considered in this study, as

$$\mathrm{SF}(y) \triangleq \| |y(t)|^2 - |y(t-l)|^2 \|. \tag{17}$$

where l is the number of delayed time-frames.

Actually, SF is a measure of spectra changing over time. It has been observed that the spectra structure of music has a higher rate of change, and goes through more drastic frame-to-frame changes than that of speech [21]. However, speech alternates between transient and non-periodic speech to short-time stationary and periodic speech due to phoneme transitions (e.g., consonant to vowel, and other phone class transitions). On the other hand, music could be periodic or monotonic and have more constant rates of change versus what is observed in speech. This indicates that the *variance* of SF (VSF) of speech, as defined (18) below, should be larger than that of music or most environmental sounds [13], [22]

$$\mathrm{VSF}(y) = E\Big\{ \big| |y(t)|^2 - |y(t-l)|^2 \big|^2 \Big\} - E\Big\{ \big| |y(t)|^2 - |y(t-l)|^2 \big| \Big\}^2. \tag{18}$$

B. Analysis of Variance of Spectral Flux

As stated above, several studies had observed larger VSF in speech compared to music or other environmental sounds [13], [22]. We are interested in revealing the reason behind. Direct analysis of VSF is difficult; but its lower bound, denoted as VSF-LB, can be obtained by (see Appendix A)

$$\mathrm{VSF}(y) \geq \text{VSF-LB}(y) = 2\Big(E\{|y(t)|^4\} - 2(E\{|y(t)|^2\})^2\Big) - 2E\{|y(t)|^2 |y(t-l)|^2\}. \tag{19}$$

Some observations can be drawn from (19): $E\{|y(t)|^4\} - 2(E\{|y(t)|^2\})^2$ is the measure of the Kurtosis, as shown in (14), and $E\{|y(t)|^2|y(t-l)|^2\}$ measures the correlation of the squared spectra across l time-frame, denoted as $\mathrm{C}(y,l) = E\{|y(t)|^2|y(t-l)|^2\}$. Therefore, VSF-LB is a compound measure of the non-Gaussianity and the correlation of the squared spectra. Fig. 2(a) shows a relationship among the Kurtosis, correlation and resulting VSF-LB, calculated from a 5-min-long speech and music recordings with normalized power in each frequency-bin. Generally, speech and music have a super-Gaussian distribution with large positive Kurtosis. Speech tends to have a larger Kurtosis than that of the music; however, this is not guaranteed in every frequency-bin. It is also interesting to look at the correlation term as shown in Fig. 2(b). When l is small, the squared spectra of speech has a higher across time correlation than that of music. However, this correlation dramatically drops to 1 as the time-frame difference l increases, which means that the squared spectra of $|y(t)|^2$ and $|y(t-l)|^2$ becomes independent for speech, i.e., $E\{|y(t)|^2|y(t-l)|^2\} = E\{|y(t)|^2\}E\{|y(t-l)|^2\} = 1$. On the other hand, although the correlation term of music is smaller when l is small, it drops more slowly. This distinct behavior verifies the observation about VSF as stated before: speech has larger frame-to-frame variability but music has a more constant rates of changes.

As the co-effect of the Kurtosis and correlation terms [shown the bottom of Fig. 2(a)], speech tends to have a larger VSF-LB than that of music, when the time difference l is sufficient large. Therefore, VSF-LB gives a new insight to interpret the observation that speech has a larger VSF than that of music.

C. Maximal VSF-Based Blind Beamforming Approach

For the task of blindly extraction of speech from a noisy environment, a cost function is needed to measure a certain desired property of speech at the output of the beamformer. Kurtosis is one of such measures that is successfully employed for a Gaussian noise environment. However, for a music environment, we need a more sophisticated measure. Inspired by the previous section, VSF could be employed as one potential speech quality measure; specifically, the larger the VSF is obtained, the more likely speech is extracted instead of music. Moreover, maximization of the VSF at the beamformer's output should lead to correct extraction of speech. The maximization of VSF can be achieved by maximizing its lower bound

VSF-LB. We define a more general cost function based on VSF-LB, namely GVSF, for the beamformer output $y(t)$ as

$$J_{\mathrm{GVSF}} = E\{|y(t)|^4\} - \alpha\big(E\{|y(t)|^2|y(t-l)|^2\} - 1\big) \quad (20)$$

where α is a scale factor that controls the balance between Kurtosis and correlation terms, as in (19). Henceforth, the optimal weights $\boldsymbol{r}$ can be found by

$$\boldsymbol{r}_{\mathrm{GVSF}} = \arg\max_{\boldsymbol{r}^H\boldsymbol{r}=1} J_{\mathrm{GVSF}}. \quad (21)$$

Notice that, the maximization of GVSF criterion is within the ICA framework but differs with an extra correlation term, which is minimized (hereby, GVSF is maximized) upon achieving independent across time-frames squared spectra. Moreover, for a source signal with temporally independent squared spectra, i.e., $E\{|y(t)|^2|y(t-l)|^2\} = E\{|y(t)|^2\}^2$, the cost function of (20) equates to $E\{|y(t)|^4)\} - \alpha E\{|y(t)|^2\}^2 - \alpha$, which differs only a constant term from the cost function proposed in [23], namely KSICA, aiming at solving the divergence problem of complex FastICA (cFastICA) [19] in the presence of Gaussian interferences.

To maximize the cost function in (21), we derive fixed-point optimization algorithms. Notice that the function J_{GVSF} can be rewritten as

$$J_{\mathrm{GVSF}} = E\{(y(t)y^*(t))^2\} - \alpha E\{y(t)y^*(t)y(t-l)y^*(t-l) - 1\} \quad (22)$$

that is, Brandwood's analyticity condition [17], [24] is satisfied. By evaluating the gradient of function J_{GVSF}, we can directly compute the derivatives with respect to the complex argument, rather than calculating individual real-valued gradients, as done in [19], which thus avoids having to deal with unnecessarily complicated expressions. For the complex derivative [24] and noting that $y = \boldsymbol{r}^H\boldsymbol{z}$, we have

$$\frac{\partial}{\partial \boldsymbol{r}^*}|y|^2 = \frac{\partial}{\partial \boldsymbol{r}^*} y^* y = y^* \boldsymbol{z}. \quad (23)$$

The gradient of the function of J_{GVSF} can be calculated as

$$\begin{aligned}
&\nabla_{\boldsymbol{r}} J_{\mathrm{GVSF}} \\
&= \frac{\partial}{\partial \boldsymbol{r}^*} J_{\mathrm{GVSF}} \\
&= \frac{\partial}{\partial \boldsymbol{r}^*} E\{|y(t)|^4\} - \alpha \frac{\partial}{\partial \boldsymbol{r}^*} \mathrm{C}(y,l) \\
&= 2E\{|y(t)|^2 y^*(t)\boldsymbol{z}(t)\} \\
&\quad - \alpha E\Big\{|y(t)|^2 y^*(t-l)\boldsymbol{z}(t-l) + |y(t-l)|^2 y^*(t)\boldsymbol{z}(t)\Big\}.
\end{aligned} \quad (24)$$

where $\mathrm{C}(y,l) = E\{|y(t)|^2|y(t-l)|^2\}$ as defined before. The fixed-point update rule can be derived in the context of constrained optimization [16], [19]. According to the Karush–Kuhn–Tucker (KKT) conditions [25], the optima of J_{GVSF} under the constraint that $\boldsymbol{r}^H\boldsymbol{r} = 1$ are obtained at points where

$$\nabla_{\boldsymbol{r}} J_{\mathrm{GVSF}} + \lambda \boldsymbol{r} = 0. \quad (25)$$

Taking the inner product of (25) with $\boldsymbol{r}$ and using (24) and constraint $\boldsymbol{r}^H\boldsymbol{r} = 1$, we have

$$2J_{\mathrm{GVSF}} - 2\alpha + \lambda = 0 \quad (26)$$

which shows that λ is real. That is, at the stable point, the gradient $\nabla_{\boldsymbol{r}} J_{\mathrm{GVSF}}$ must be equal to $\boldsymbol{r}$ multiplied by some scalar constant. For such a case, adding the gradient to $\boldsymbol{r}$ does not change its direction, and therefore achieve convergence, allowing us to optimize the objective using the fixed-point theory. We obtain the one-unit complex fixed-point algorithm as

$$\begin{aligned}
\nabla_{\boldsymbol{r}} J_{\mathrm{GVSF}} &\rightarrow \boldsymbol{r} \\
\frac{\boldsymbol{r}}{\|\boldsymbol{r}\|} &\rightarrow \boldsymbol{r}.
\end{aligned} \quad (27)$$

Note that convergence of the fixed-point algorithm implies that the old and new values of $\boldsymbol{r}$ must point in the same direction.

D. Stability Analysis

In this section, stability conditions for the GVSF algorithm are analyzed. In Appendix B, using a second-order Taylor expansion of the cost function around the stability point, we show that the local maximum of J_{GVSF} is achieved for a given source, denoted as s_j ($j \leq M$), when the following condition is satisfied:

$$\frac{\psi(s_j)}{2} \triangleq E\{|s_j(t)|^4\} - \alpha E\{|s_j(t)|^2|s_j(t-l)|^2\} > 0. \quad (28)$$

Some observations can be drawn here. First, let us set $\alpha = 2$. From Cauchy–Schwarz inequality and noticing that $E\{|s_j(t)|^4\} = E\{|s_j(t-l)|^4\}$, we obtain

$$E\{|s_j(t)|^4\} \geq E\{|s_j(t)|^2|s_j(t-l)|^2\} \quad (29)$$

with equalization when $|s_j(t)|^2$ and $|s_j(t-l)|^2$ are correlated. Therefore, for a source with a highly correlated squared spectra between two time frames, (29) approximately holds and $\psi(s_j)/2 \approx -E\{|s_j(t)|^4\}$ is less than or equal to zero. In such a condition, the proposed algorithm hardly converges and fails to extract such sources. On the other hand, the speech squared spectra will generally has a larger variation when compared to other environmental sounds and is approximately independent. Therefore, we can approximate $\psi(s_j)/2 \approx E\{|s_j(t)|^4\} - 2E\{|s_j(t)|^2\}^2 = \mathrm{K}(s_j) > 0$ for speech with a super-Gaussian distribution, i.e., the proposed GVSF algorithm could converge and thus extract the speech.

E. Adaptive α Design

For real world applications, a suitable α which controls the balance between the Kurtosis term and the correlation term and thus satisfies the convergence condition of speech as well as the non-convergence condition of music, is hard to pre-chosen and fixed in real-time process, since the Kurtosis is sensitive to outliers [16] and generally has a large dynamic range. However, speech tends to have independent squared spectra when the time-frames difference l increases; that is, the relation $\mathrm{C}(y,l) - 1 = 0$ holds when l is greater than some particular value. Therefore, for extraction of speech rather than music, the cost function

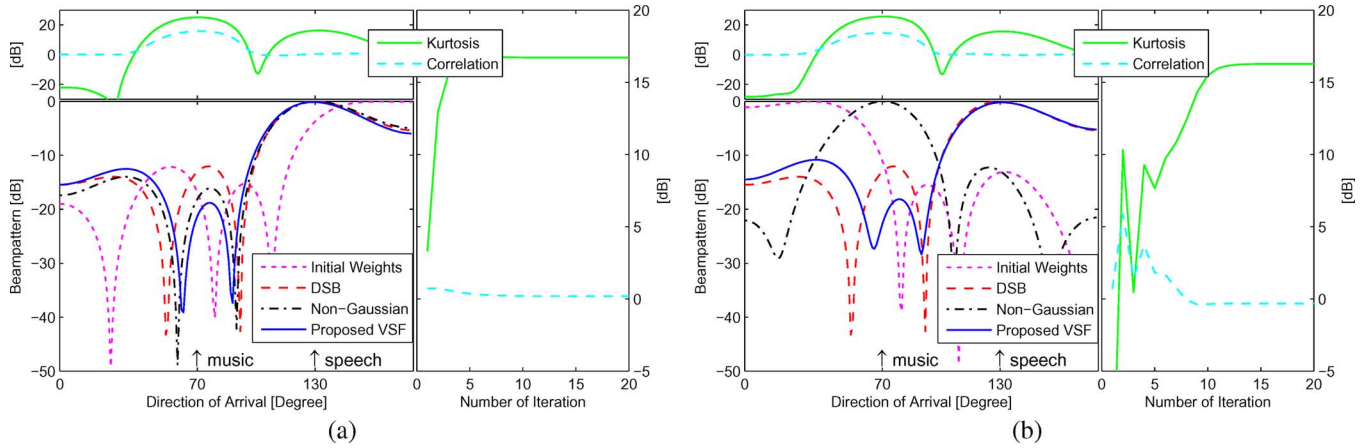

Fig. 3. Illustration the performance of the proposed method. (a) Non-Gaussianity maximization leads to reinforcement at the speech DOA (130°), if initial weights are properly chosen. (b) Non-Gaussianity maximization leads to reinforcement at the music DOA (70°), if initial weights are closely pointing towards the noise. However, the proposed approach can robustly enhance the speech while suppressing the noise for both of the situations. Left-top: spatial distribution of the Kurtosis (solid) and correlation (dash). Left-bottom: beampattern formed by: Delay-and-sum beamformer (DSB) using actual array steering vector of the desired speech (dash), initial weights for both of the non-Gaussian maximization and the proposed approach (dot), beamforming weights of the non-Gaussian maximization (dash-dot), and beamforming weights of the proposed approach (solid). Right: Measured output Kurtosis (solid) and correlation (dash) at each iteration of proposed method.

of (20) can be reformulated as the Lagrange method for a constrained maximization problem, as

$$\text{maximize } E\{|y(t)|^4\} \quad \text{s.t. } \mathrm{C}(y,l)-1=0 \tag{30}$$

and hence, α can be treated as a Lagrange multiplier, which can be adaptively updated by

$$\alpha=\beta\big(\mathrm{C}(y,l)-1\big) \tag{31}$$

where β is a nonsensitive pre-chosen positive scalar.

F. Illustrative Example

Fig. 3 illustrates the effectiveness of the proposed approach. One speech source and one music interference impinge on a five-element 4.5-cm uniformly spaced array from DOAs of 130° and 70°, respectively. The frequency is approximately 2.5 kHz, the signal-to-interference ratio (SIR) is adjusted to be 0 dB, and signal-to-background spatial white noise is set to 5 dB. As shown, the beamformer weights formed by Non-Gaussianity criterion as in (16) may strengthen the signal from either one of directions of the two signals while suppressing the other, greatly depending on the initial weights. The proposed approach has the ability to conquer the ambiguity problem, resulting in the successful enhancement towards the speech DOA while suppressing the music interference.

V. Implementation Issue and Batch Processing Algorithm Summary

A. Combating Divergence

The normalized power assumption of the two terms $E\{|y(t)|^2\}=1$ and $E\{|y(t-l)|^2\}=1$ is used for the optimization steps throughout the derivation in the previous section. However, as pointed out in [17] and [23], $E\{|y(t)|^2\}$ is a function of $\boldsymbol{r}$ and thus not a constant without the additional normalization step on $\boldsymbol{r}$. Hence, the directions of these two terms should be considered before the normalization step, and better results can be expected. Therefore, the gradient of the full Kurtosis can be obtained as

$$\frac{\partial \mathrm{K}(y)}{\partial \boldsymbol{r}^*}=2\Big(E\{|y(t)|^2y^*(t)\boldsymbol{z}(t)\}-2E\{|y(t)|^2\}E\{y^*(t)\boldsymbol{z}(t)\}\Big). \tag{32}$$

For the correlation term $\mathrm{C}(y,l)=E\{|y(t)|^2|y(t-l)|^2\}$, we consider the normalized correlation as

$$\bar{\mathrm{C}}(y,l)=\frac{E\{|y(t)|^2|y(t-l)|^2\}}{E\{|y(t)|^2\}E\{|y(t-l)|^2\}}=\frac{E\{|y(t)|^2|y(t-l)|^2\}}{E\{|y(t)|^2\}^2}. \tag{33}$$

and its gradient can be derived as

$$\frac{\partial \bar{\mathrm{C}}(y,l)}{\partial \boldsymbol{r}^*}=\frac{\frac{\partial \mathrm{C}}{\partial \boldsymbol{r}^*}E\{|y(t)|^2\}^2-2E\{|y(t)|^2\}E\{y^*(t)\boldsymbol{z}(t)\}\mathrm{C}(y,l)}{E\{|y(t)|^2\}^4}. \tag{34}$$

where

$$\frac{\partial \mathrm{C}(y,l)}{\partial \boldsymbol{r}^*}=E\Big\{|y(t)|^2y^*(t-l)\boldsymbol{z}(t-l)+|y(t-l)|^2y^*(t)\boldsymbol{z}(t)\Big\}. \tag{35}$$

Therefore, considering the adaptive balancing term α in (31), the final non-diverging gradient becomes

$$\nabla_{\boldsymbol{r}}\tilde{J}_{\mathrm{GVSF}}=\frac{\partial \mathrm{K}(y)}{\partial \boldsymbol{r}^*}-\beta(\bar{\mathrm{C}}(y,l)-1)\frac{\partial \bar{\mathrm{C}}(y,l)}{\partial \boldsymbol{r}^*}. \tag{36}$$

B. Algorithm Summary

We implement our proposed algorithm for the task of speech enhancement. A Batch processing version is summarized in this section. Suppose the input array observation data is collected for a certain amount of time, corresponding to an overall batch

duration of T time-frames. Here, we perform EVD of the covariance matrix of the array observation data, which henceforth can be projected into the signal subspace, as follows.

1) Estimate covariance matrix from T time frames:

$$\hat{R}_{xx} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}(t)\boldsymbol{x}^H(t).$$

2) Perform EVD of $\hat{R}_{xx}$, form the signal subspace U_s using the eigenvector corresponding to the J largest eigenvalue

$$\{U_s, \Lambda_s\} \leftarrow \mathrm{EVD}(\hat{R}_{xx})$$

and construct the normalized subspace

$$\bar{U}_s = [\lambda_1^{-1/2}\boldsymbol{u}_1, \lambda_2^{-1/2}\boldsymbol{u}_2, \ldots \lambda_J^{-1/2}\boldsymbol{u}_J].$$

3) Project data into the normalized signal subspace, resulting in dimension reduced data

$$\boldsymbol{z}(t) = \bar{U}_s^H \boldsymbol{x}(t).$$

With this, the steps below follow an iterative schedule for each data batch in order to numerically find an optimal solution.

4) Initialize the rotation vector $\boldsymbol{r}$, and compute the *a priori* output signal:

$$y(t) = \boldsymbol{r}^H \boldsymbol{z}(t).$$

5) Update the rotation vector as

$$\boldsymbol{r} \leftarrow \nabla_{\boldsymbol{r}} \tilde{J}_{\mathrm{GVSF}} \quad \text{as in Eq.(36)}$$
$$\boldsymbol{r} \leftarrow \frac{\boldsymbol{r}}{\|\boldsymbol{r}\|}.$$

6) When a stopping criterion is meet, or if a specific number of iterations has passed, let $\boldsymbol{r}_o = \boldsymbol{r}$, and stop the iterations for this batch.

7) Output the enhanced speech as

$$y(t) = (\bar{U}_s \boldsymbol{r}_o)^H \boldsymbol{x}(t),$$
$$y(t) \leftarrow |y(t)| \exp\{-j \cdot \mathrm{Phase}(x_1(t))\}.$$

In the last step, the phase from a certain channel (e.g., the first channel) of the array observation data is used to reconstruct the enhanced speech, instead of directly using the phase of the enhanced signal. The reason is that upon convergence, the proposed algorithm may introduce random phase shifts for each frequency-bin. Processed speech sounds more natural by using the unprocessed phase, the same procedure conventionally employed for single channel speech enhancement [18].

VI. Evaluation

We evaluate the newly derived blind beamforming algorithm for two distinct scenarios: 1) diffuse music noise field, and 2) directional music interference noise field. All microphone recordings are sampled at 24 kHz and downsampled to 8 kHz and quantized to 16 bits for processing. The framewise spectra analysis uses a Hamming window of length $\mathrm{L} = 256$ samples, with

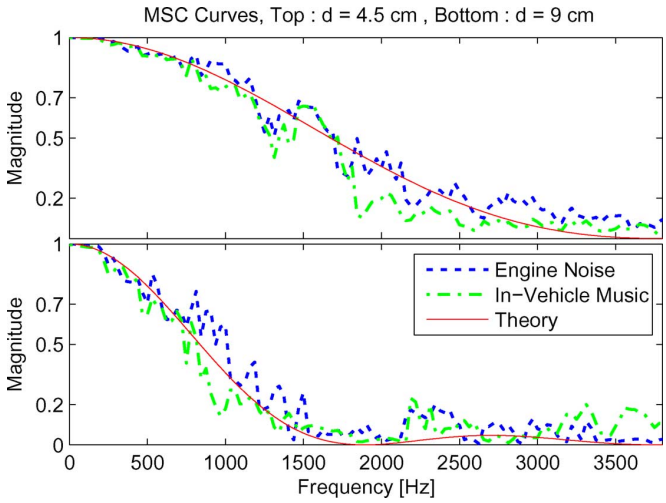

Fig. 4. Measured MSC curves for real in-vehicle noise.

128 sample overlapping frames (FFT length $\mathrm{L} = 256$). A rectangular window is used for overlap-add of the enhanced frames.

A. Evaluation in Diffuse Music Noise Field

1) Diffuse Noise Field: A diffuse or spherically isotropic noise field can serve as an effective model for many applications concerning practical reverberant noise environments encountered in speech enhancement applications, such as offices and cars [26]. A generally accepted characterization of a diffuse field is one where noise sources have equal power propagating in all directions simultaneously. It can be shown that the covariance matrix R_{nn} of a diffuse noise field is real-valued with entries in the ith row and jth column is given by [26]

$$\{R_{nn}\}_{i,j} = \sigma_n^2 \Gamma_{i,j} \tag{37}$$

with

$$\Gamma_{i,j} = \mathrm{sinc}\left(\frac{2\pi d_{ij} f}{c}\right) \tag{38}$$

where d_{ij} is the distance between the ith and jth microphone, f is the frequency and c is the speed of sound (e.g., 340 m/s).

As a verification, the magnitude squared coherence function (MSC)

$$\mathrm{MSC}_{i,j}(f) = \Gamma_{i,j}^2(f) \tag{39}$$

is analyzed for UTDrive [27] in-vehicle noise database. In Fig. 4, MSC curves are plotted for the recordings when only the vehicle engine is stably working (i.e., engine noise) and when the audio system is also turned on (i.e., music noise), respectively. Since the audio system has multiple speakers, the in-vehicle music also tend to be diffuse. As can be observed, the measured MSC curves closely follow their theoretical values.

Unlike a spatially white noise field with an identity covariance matrix, whose eigenvalues are uniformly distributed in the entire eigenspace, the eigenvalues of the diffuse noise covariance matrix are concentrated along only a few eigenvectors. Next, we show that this compact eigenspace can lead to a biased or ambiguous DOA identification problem, if the second-order

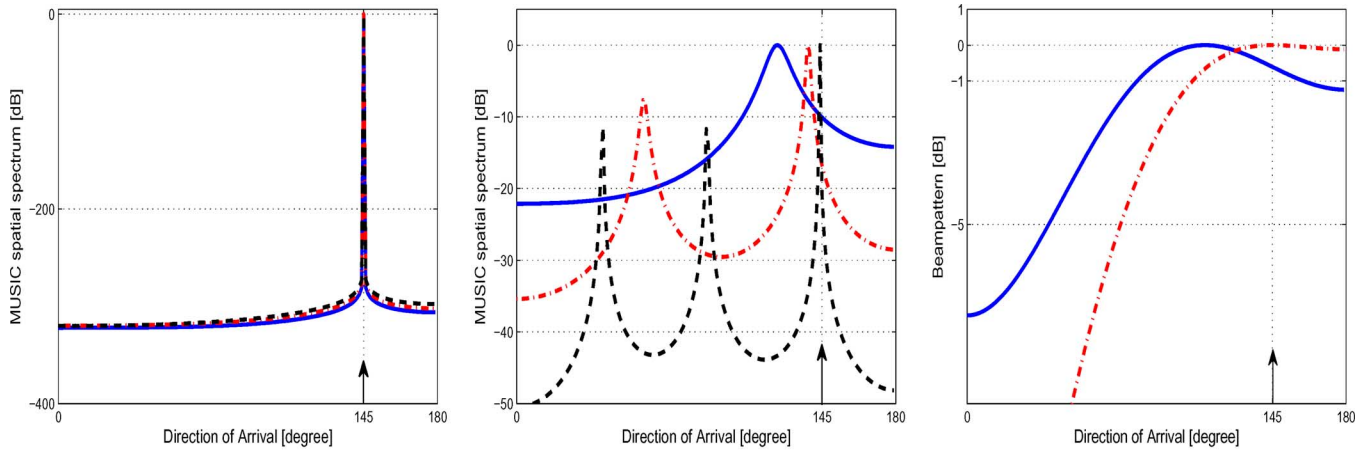

Fig. 5. Illustration of the array steering vector/DOA bias and ambiguity using SOS-based approaches for MUSIC-based DOA identification using eigenvectors corresponding to the smallest first eigenvalue (dash), smallest two eigenvalues (dash dot) and least three eigenvalues (solid), in (a) spatially white noise and (b) diffuse noise fields, respectively. (c) Beampattern of the principle eigenvector in spatially white (dash) and diffuse (solid) noise fields.

statistics (SOS)-based subspace approaches, such as MUSIC [14] or principle eigenvector beamformer are employed.

In Fig. 5, we try to identify the DOA of a target signal at frequency of 1 kHz by a four-element uniformly linearly spaced array (ULA) with 4.5 cm between consecutive sensors. As shown in Fig. 5(a), in a spatially white noise field, MUSIC approach can accurately form a unique peak towards the target DOA (145°), independent of the noise spaces (null spaces) that are constructed by any combinations of the eigenvectors corresponding to the smallest 1, 2, or 3 eigenvalues. However, for the case of diffuse noise, as shown in Fig. 5(b), a DOA bias is introduced if the noise space is formed by the eigenvectors corresponding to the three smallest eigenvalues; multiple peaks and also bias of DOA are formed if the noise space is constructed by any combination of the eigenvectors corresponding to the smallest one or two eigenvalues. Similarly, the beampattern formed by the principle eigenvector is also biased in the diffuse noise field, as compared in Fig. 5(c).

In fact, SOS-based DOA identification approaches search for the spatial power extremes, through a process of scanning over a certain space (e.g., noise subspace or signal subspace). When the diffuse noise field is taken into consideration, multiple power extremes are present, resulting in a DOA either ambiguity or bias problem. HOS-based DOA identification approaches [5] can solve this problem if the target is non-Gaussian while the noise field is Gaussian distributed, since the HOS of a Gaussian distributed noise is vanished. However, this method may fail in microphone array-based applications, due to the fact that acoustic noises (e.g., music noises) generally possess a super-Gaussian probability distribution.

2) Evaluation With In-Vehicle Diffuse Music Noise Field: The topic of capturing clean and distortion-free speech under distant talker conditions using a microphone array within noisy in-car environments has attracted much attention [1], [28]. Next, we show the effectiveness of our proposed algorithm in a real in-vehicle diffuse music noise field, which can be regarded as a super-Gaussian distributed diffuse noise field. The UTDrive corpus [27] is used for the evaluation. The microphone array constructed for the UTDrive project is a linear 5-element array,

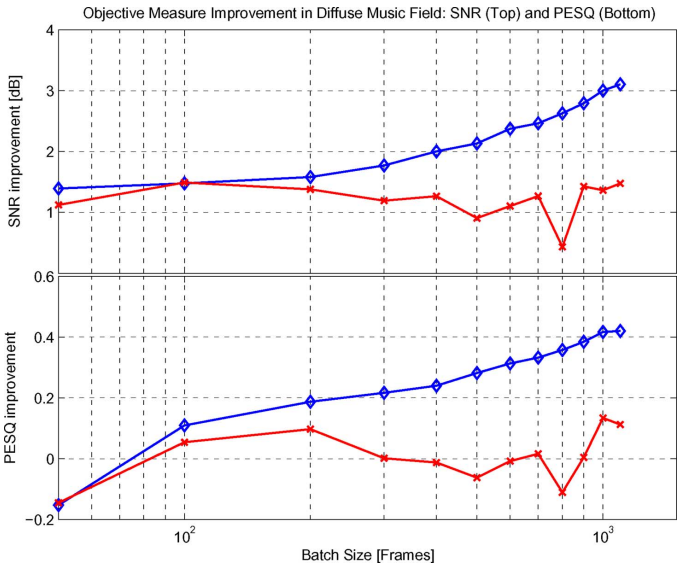

Fig. 6. Performance of the proposed approach (◇) and KSICA (×) with increased batch size for in-vehicle speech enhancement. Performance improvement in terms of SNR (Top) and PESQ (Bottom) over the observation data from a single microphone.

with 4.5 cm spacing between consecutive microphones to avoid spatial aliasing, given a frequency bandwidth of 4 kHz. The array is mounted on the top of the dashboard in front of the passenger seat, approximated 120° from the driver's seat. In addition, a close talk microphone is also used as a clean speech reference. The recordings used for this evaluation is collected in a Toyota RAV 4WD driving stably on a highway (approximately 90 km/h) with all windows closed and the audio system was turned on for music noise generation.

As a comparison with our proposed algorithm, the KSICA [23] speech extraction algorithm, an improved version of the cFastICA [19], is also evaluated. Because the performance of these two algorithms are greatly dependent on their initial weights, we use the same initial weights for all the algorithms at each processing data batch, but randomly generate new entry values for every batch. For different batch sizes, Fig. 6

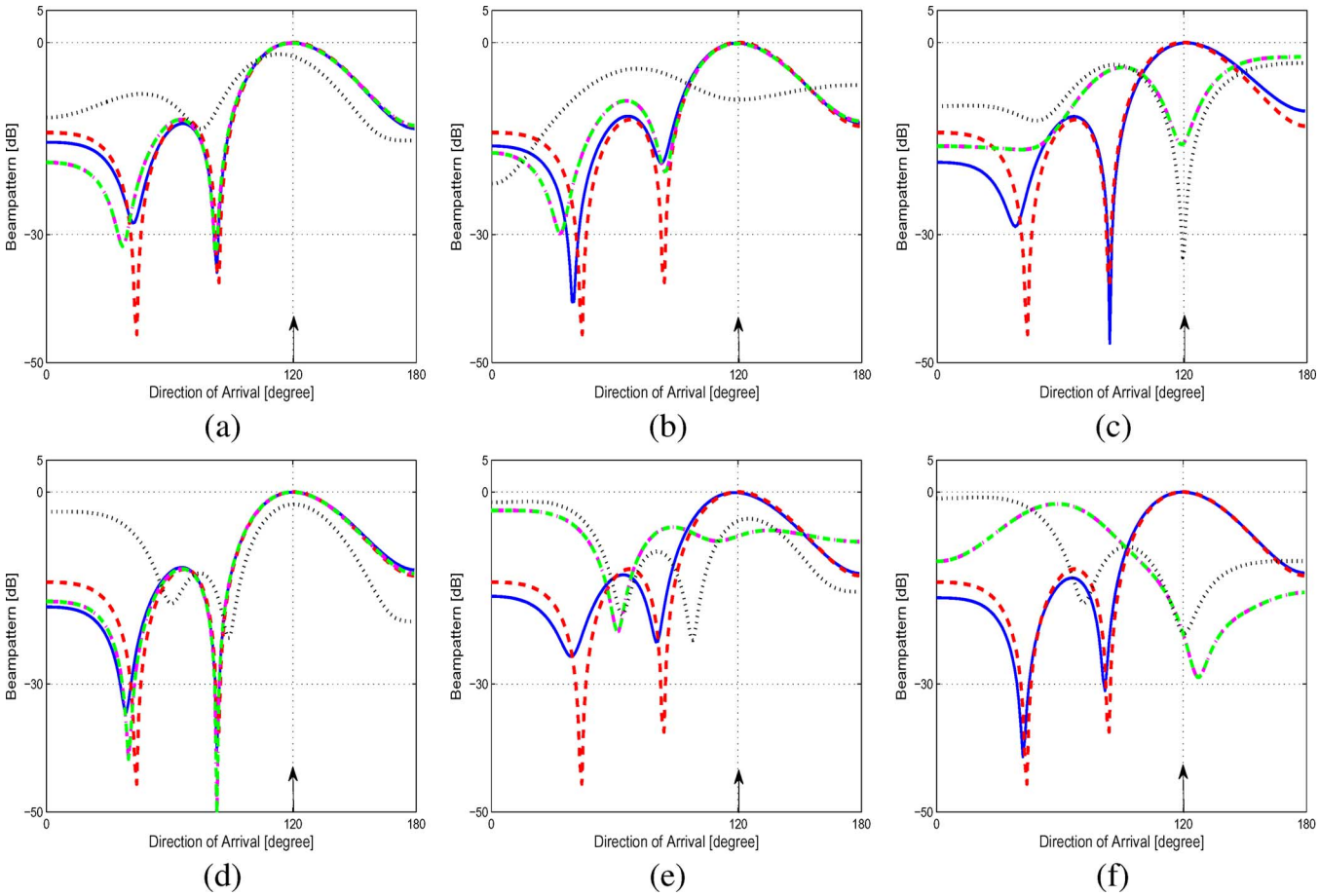

Fig. 7. Beampatterns generated by different algorithms in the real in-vehicle noisy environment: actual array steering vector (dashed line), initial weights (dot line), cFastICA and KSICA (dashed-dot line, these two algorithms result in highly overlapped beampatterns) and the proposed approach (solid line). (a)-(c), beampatterns are formed when noise is approximately Gaussian distributed (e.g., engine noise dominates). (e)-(f), beampatterns are formed when the noise is approximately super-Gaussian distributed (e.g., music dominates).

presents the speech quality (SNR and PESQ [29]) improvement measured at the beamformer outputs over one single channel observation, using close-talk microphone as a reference. In a diffuse noise field, the speech quality improvement provided by the microphone array is not significant, as shown in Fig. 6. Proposed algorithm has the ability to achieve higher speech quality improvements by correctly identifying the array steering vector; while KSICA sometimes performs even worse than the original input for this real diffuse noise field.

Fig. 7 shows same typical beampatterns at the frequency of approximately 2.5 kHz, obtained by the proposed algorithm and the KSICA. When the in-vehicle noise is approximately Gaussian distributed (e.g., the scenario that the engine noise (Gaussian distributed) dominates the noise field), KSICA is capable of converging to the weights that are maximally strengthening towards the speech DOA, as shown in Fig. 7(a) and (b), except for the cases of poorly initialized weights, as seen in Fig. 7(c). However, when the noise distribution becomes super-Gaussian (e.g., the scenario that the music (super-Gaussian distributed) dominates the noise field), beamforming weights generated by KSICA is rather random, except for the cases that the initial weights are close to the optimal one. In all the conditions, the proposed approach robustly achieves the optimal weights, thereby providing constant enhancement for the desired speech.

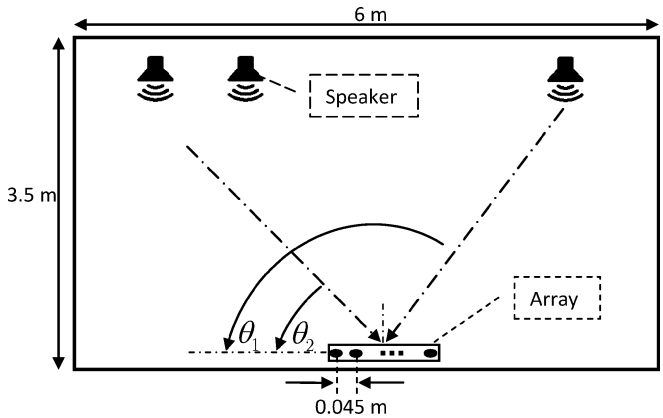

Fig. 8. Overhead view of simulation environment.

B. Evaluation With Directional Music Interference

1) Simulation Setup: In this section, the performance of the proposed approach is evaluated when the music interferences are simultaneously present with the desired speech. The experiment was conducted in a moderately reverberant room ($T_{30} \approx 50$ ms), and all recordings employed a linear array of cardioid condenser microphones with spacing between consecutive microphones of 4.5 cm. The number of microphones used was

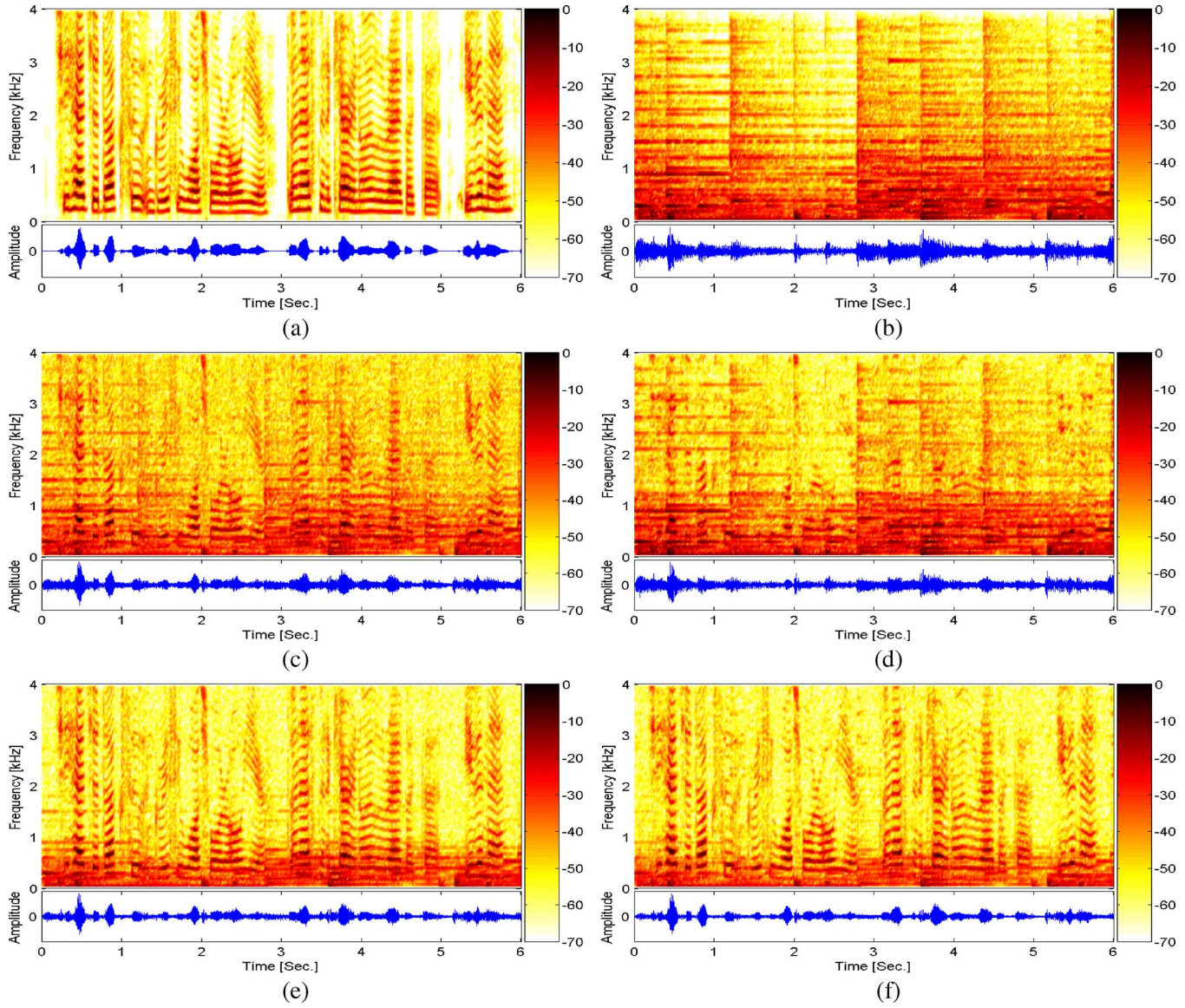

Fig. 9. Spectrograms (top) [in dB scale (right)] and time-domain waveforms (bottom), for (a) clean speech signal; (b) music interference; (c) one-channel observation data: speech and music interference project on the array with DOAs 135° and 45°, respectively; SIR (speech to music) is 0 dB and SNR (speech to spatially white noise) is 5 dB; (d) output using non-Gaussianity maximization (cFastICA); (e) output of DSB with perfect knowledge of the array steering vector for the desired speech; (f) output of the proposed approach. (a) Clean speech. (b) Music interference. (c) Single-channel observation. (d) non-Gaussianity criterion (cFastICA). (e) Optimal delay-and-sum beamformer. (f) Proposed method.

varied from two to eight and the complete array was previously calibrated. The locations/DOAs of the multiple sources were pre-selected and fixed, according to different experimental setups. The speech data was chosen from the TIMIT database [30] and playback using positioned speakers as shown in Fig. 8. For each evaluation, in total of 20 minutes recorded data set was adopted.

2) Evaluation With Music Interference: As presented in Section IV-F, the non-Gaussianity criterion-based BSE approach possesses an ambiguity problem as long as the non-Gaussian interferences are present. Fig. 9 shows a series of the array output spectrograms, using the proposed approach compared with cFastICA algorithm with randomly initialized weights. Approximately 6 s of speech utterance and music interference were projected onto the array with DOAs of 135° and 45° respectively, with an overall SIR (speech-to-music) of 0 dB and an SNR (speech to spatially white noise) of 5 dB. As shown, the non-Gaussianity-based methods fail to extract the speech properly. However, the proposed method outperforms even the delay-and-sum beamfomer (DSB) which has perfect knowledge of the array steering vector of the speech, since the proposed approach can not only enhance towards the speech DOA, but also has the ability to suppresses towards the DOA of music.

Fig. 10 directly compares the measured gains for the desired speech and music interference. It should be noted that, the proposed approach has its own weakness and did fail to extract the speech at a couple of frequency-bins, where the VSF of speech is quite similar to that of the music. Thus, maximization of the output VSF may not always result in the optimal weights that

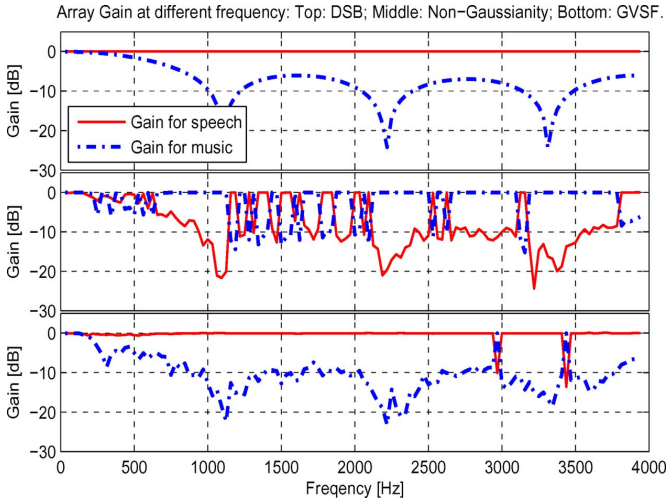

Fig. 10. Array gain for the speech (solid) and music (dash-dot) for different approaches: top: optimal DSB; middle: non-Gaussianity; bottom: proposed GVSF.

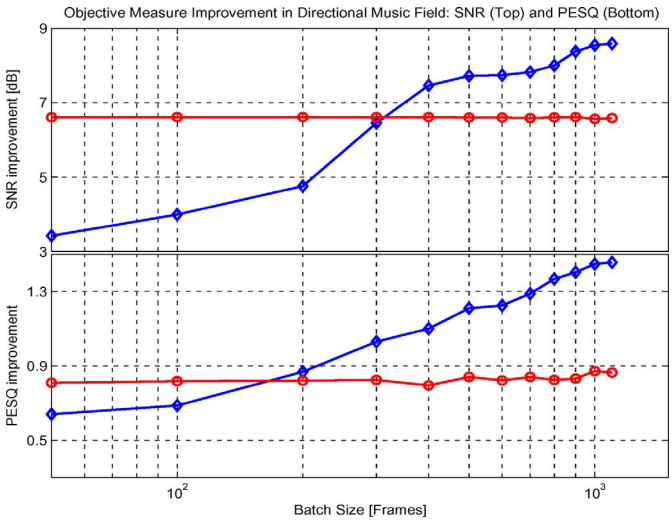

Fig. 11. Performance of the proposed approach (◇) with increased batch sizes for a five-element array. The performance of DSB (○) with perfect knowledge of the desired speech steering vector is also drawn for comparison. Performance improvement in terms of SNR (top) and PESQ (bottom) over observation data from a single microphone are shown.

places maximal gain towards the speech DOA but least gain towards the DOA of music interference. However, this problem could be alleviated through a combination of DOA decisions of each frequency.

Next, the objective measure of speech enhancement performance is studied for different processing batch sizes, as shown in Fig. 11 for a five-element array. The "worse case" initial weights (i.e., initialized with the array steering vector of the interference), are employed and the performance of DSB with perfect knowledge of the speech steering vector is compared. With increased batch size, the proposed approach not only achieves successful identification of the speech DOA, it also has less array gain towards the music DOA, resulting in better performance versus DSB. Finally, the performance of cFastICA is omitted here since it tends to enhance the noise while nulling out the desired speech.

VII. Conclusion and Discussion

A new promising time–frequency domain blind beamforming approach for extracting the desired speech from a noisy music environment has been presented, along with the evaluations conducted in both real in-vehicle and simulated noisy environments. Compared to the conventional BSE approaches that rely on various *a priori* information, the proposed approach is totally blind. The proposed approach is based on the concept that through the maximization of a particular measure of the discriminative speech features, speech can be extracted without ambiguity. We have shown that the VSF can serve as one of such effective features. An algorithm was developed that is focused on the maximization of the VSF criterion, including a complex-value fixed-point optimization process. In the evaluation, the effectiveness of the proposed approach is confirmed using real UTDrive in-vehicle recordings, as well as a simulated recordings. The results show consistent improvement of our proposed approach over traditional methods.

While advancement has been achieved, robust BSE or blind beamforming for the speech remains a challenging problem. As absolved in our experiments, the proposed blind algorithm still may fail to extract the desired speech at sometimes. This is particularly true when the processing batch is not large enough, or the VSF of speech is not sufficiently discriminative from the environmental music interferences. Further research into enhanced speech features and criteria is an ongoing challenge.

The following Appendix presents derivations for a lower bound of VSF and stable conditions for the proposed algorithm.

Appendix A
Lower Bound of VSF

In Section IV-B, the criterion for the variance of the spectral flux was presented. Here, we consider the formulation of its lower bound. From (18),

$$\mathrm{VSF} = E\left\{\left||y(t)|^2 - |y(t-l)|^2\right|^2\right\} - E\left\{\left||y(t)|^2 - |y(t-l)|^2\right|\right\}^2 \tag{40}$$

it is obvious that

$$E\left\{\left||y(t)|^2 - |y(t-l)|^2\right|\right\}^2 \le E\left\{\left||y(t)|^2 + |y(t-l)|^2\right|\right\}^2. \tag{41}$$

With this, we have the following relation:

$$\mathrm{VSF} \ge E\left\{\left||y(t)|^2 - |y(t-l)|^2\right|^2\right\} - E\left\{\left||y(t)|^2 + |y(t-l)|^2\right|\right\}^2. \tag{42}$$

Note that, $E\{|y(t)|^4\} = E\{|y(t-l)|^4\}$ and $E\{|y(t)|^2\} = E\{|y(t-l)|^2\}$, and therefore the right hand side (RHS) of (42) can be simplified as

$$\mathrm{RHS} = 2\Big(E\{|y(t)|^4\} - 2\big(E\{|y(t)|^2\}^2\big)\Big) - 2E\{|y(t)|^2|y(t-l)|^2\}. \tag{43}$$

Some observations can now be drawn from (43): $E\{|y(t)|^4\} - 2(E\{|y(t)|^2\})^2$ is the measurement of the Kurtosis for a complex circular distributed variable and $E\{|y(t)|^2|y(t-l)|^2\}$ measures the correlation of the squared spectra between two time frames.

APPENDIX B
DERIVATION OF THE STABILITY CONDITIONS

The derivation of stability follows the analysis of the complex FastICA method presented in [19] and [17], and the complex derivative in [24]. Assume in total of J sources, $s_1(t), s_2(t), \ldots, s_J(t)$, impinge on the microphone array with array steering vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_J$, respectively; denote $\boldsymbol{s}(t) = [s_1(t), s_2(t), \ldots, s_J(t)]^T$ and $A = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_J]$, then we get the observation data vector $\boldsymbol{x}(t) = A\boldsymbol{s}(t)$. By defining a linear transform $\boldsymbol{q} = A^H \bar{U}_s \boldsymbol{r}$, the transformed cost function of (20) can be written as

$$J_{\mathrm{GVSF}}(\boldsymbol{q}) = E\{|\boldsymbol{q}^H \boldsymbol{s}(t)|^4\} - \alpha\Big(E\{|\boldsymbol{q}^H \boldsymbol{s}(t)|^2 |\boldsymbol{q}^H \boldsymbol{s}(t-l)|^2\} - 1\Big). \tag{44}$$

The stability analysis is conducted by evaluating a second-order Taylor series expansion around the optimal point $\boldsymbol{q}_o$ that extracts one desired source, e.g., s_1, without loss of generality, while rejecting the others. That is, $\boldsymbol{q}_o = A^H \bar{U}_s \boldsymbol{r}_o = [\delta, 0, \ldots, 0]^T$, with $|\delta| = 1$, and thus $y(t) = \delta^* s_1(t)$ with only a phase difference compared to the desired source s_1.

Define $\boldsymbol{e} = [1, 0, \ldots, 0]^T$ and $\psi(s_1) = 2E\{|s_1(t)|^4\} - 2\alpha E\{|s_1(t)|^2 |s_1(t-l)|^2\}$, the two first-order gradients of $J_{\mathrm{GVSF}}(\boldsymbol{q})$ evaluated at $\boldsymbol{q}_o$ can be obtained as

$$\frac{\partial J_{\mathrm{GVSF}}(\boldsymbol{q}_o)}{\partial \boldsymbol{q}} = [\psi(s_1)\delta^*]\boldsymbol{e} \tag{45}$$

and

$$\frac{\partial J_{\mathrm{GVSF}}(\boldsymbol{q}_o)}{\partial \boldsymbol{q}^*} = [\psi(s_1)\delta]\boldsymbol{e}. \tag{46}$$

Evaluating the four Hessians at the point $\boldsymbol{q}_o$ and using the condition of independence between elements of $\boldsymbol{s}$, i.e., $s_1, s_2, \ldots, s_J$, we obtain

$$\frac{\partial J_{\mathrm{GVSF}}(\boldsymbol{q}_o)}{\partial \boldsymbol{q} \partial \boldsymbol{q}^T} = [\psi(s_1)\delta^{*2}]\boldsymbol{e}\boldsymbol{e}^T \tag{47}$$

$$\frac{\partial J_{\mathrm{GVSF}}(\boldsymbol{q}_o)}{\partial \boldsymbol{q} \partial \boldsymbol{q}^H} = [2\psi(s_1)]\boldsymbol{e}\boldsymbol{e}^T \tag{48}$$

$$\frac{\partial J_{\mathrm{GVSF}}(\boldsymbol{q}_o)}{\partial \boldsymbol{q}^* \partial \boldsymbol{q}^T} = [2\psi(s_1)]\boldsymbol{e}\boldsymbol{e}^T \tag{49}$$

and

$$\frac{\partial J_{\mathrm{GVSF}}(\boldsymbol{q}_o)}{\partial \boldsymbol{q}^* \partial \boldsymbol{q}^H} = [\psi(s_1)\delta^2]\boldsymbol{e}\boldsymbol{e}^T. \tag{50}$$

Introducing a small perturbation $\triangle\boldsymbol{q} = [\triangle q_1, \triangle q_2, \ldots, \triangle q_J]$ around point $\boldsymbol{q}_o$, we can write the second-order Taylor series expansion of $J_{\mathrm{GVSF}}(\boldsymbol{q})$ around $\boldsymbol{q}_o$ as [31]

$$\begin{aligned} &J_{\mathrm{GVSF}}(\boldsymbol{q}_o + \triangle\boldsymbol{q}) - J_{\mathrm{GVSF}}(\boldsymbol{q}_o) \\ &= \Big(\frac{\partial J_{\mathrm{GVSF}}(\boldsymbol{q}_o)}{\partial \boldsymbol{q}}\Big)^T \triangle\boldsymbol{q} + \Big(\frac{\partial J_{\mathrm{GVSF}}(\boldsymbol{q}_o)}{\partial \boldsymbol{q}^*}\Big)^T \triangle\boldsymbol{q}^* \\ &+ \frac{1}{2}\triangle\boldsymbol{q}^T \frac{\partial J_{\mathrm{GVSF}}(\boldsymbol{q}_o)}{\partial \boldsymbol{q} \partial \boldsymbol{q}^T}\triangle\boldsymbol{q} + \frac{1}{2}\triangle\boldsymbol{q}^H \frac{\partial J_{\mathrm{GVSF}}(\boldsymbol{q}_o)}{\partial \boldsymbol{q} \partial \boldsymbol{q}^H}\triangle\boldsymbol{q} \\ &+ \frac{1}{2}\triangle\boldsymbol{q}^T \frac{\partial J_{\mathrm{GVSF}}(\boldsymbol{q}_o)}{\partial \boldsymbol{q}^* \partial \boldsymbol{q}^T}\triangle\boldsymbol{q}^* + \frac{1}{2}\triangle\boldsymbol{q}^H \frac{\partial J_{\mathrm{GVSF}}(\boldsymbol{q}_o)}{\partial \boldsymbol{q}^* \partial \boldsymbol{q}^H}\triangle\boldsymbol{q}^* \\ &+ o(\|\triangle\boldsymbol{q}\|^2). \end{aligned} \tag{51}$$

Using (45)–(50), we can rewrite (51) as

$$\begin{aligned} &J_{\mathrm{GVSF}}(\boldsymbol{q}_o + \triangle\boldsymbol{q}) - J_{\mathrm{GVSF}}(\boldsymbol{q}_o) \\ &\quad = \psi(s_1)\Big(2Re\{\delta^* \triangle q_1\} + Re\{\delta^{*2} \triangle q_1^2\} + 2|\triangle q_1|^2\Big) \\ &\qquad + o(\|\triangle\boldsymbol{q}\|^2). \end{aligned} \tag{52}$$

Due to the constraint of the optimal solution $\|\boldsymbol{q}_o\| = 1$ and the permuted optimal solution $\|\boldsymbol{q}_o + \triangle\boldsymbol{q}\| = 1$, the following relationship holds [19]

$$2Re\{\delta^* \triangle q_1\} = -\|\triangle\boldsymbol{q}\|^2 \tag{53}$$

which implies that the terms of order $\triangle q_1^2$ in (52) become $o(\|\triangle\boldsymbol{q}\|^2)$, i.e., of higher order and can be neglected. This relationship yields

$$J_{\mathrm{GVSF}}(\boldsymbol{q}_o + \triangle\boldsymbol{q}) - J_{\mathrm{GVSF}}(\boldsymbol{q}_o) \approx -\psi(s_1)\|\triangle\boldsymbol{q}\|^2. \tag{54}$$

Because the term $\|\triangle\boldsymbol{q}\|^2$ is always non-negative, $\boldsymbol{q}_o$ is an local maximum of $J_{\mathrm{GVSF}}(\boldsymbol{q})$ if $\psi(s_1) > 0$.

REFERENCES

[1] M. Brandstein and D. Ward, *Microphone Arrays*. New York: Springer, 2001.

[2] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.

[3] G. Kim and N. I. N. Cho, "Frequency domain multi-channel noise reduction based on the spatial subspace decomposition and noise eigenvalue modification," *Speech Commun.*, vol. 50, pp. 382–391, Sep. 2008.

[4] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 5, pp. 1529–1539, Jul. 2007.

[5] J. F. Cardoso and A. Souloumiac, "Blind beamforming for non-Gaussian signals," *IEE Proc. Radar Signal Process.*, vol. 140, pp. 362–370, 1993.

[6] Z. Ding, "A new algorithm for automatic beamforming," in *Proc. Conf. Signals, Syst., Comput.*, 1991, vol. 2, pp. 689–693.

[7] L. Parra and C. Alvino, "Geometric source separation: merging convolutive source separation with geometric beamforming," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 352–362, Sep. 2002.

[8] Y. Takahashi, T. Takatani, K. Osakoand, H. Saruwatari, and K. Shikano, "Blind spatial subtraction array for speech enhancement in noisy environment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 650–664, May 2009.

[9] M. Knaak, S. Araki, and S. Makino, "Geometrically constrained independent component analysis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 715–726, Feb. 2007.

[10] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Leeand, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 666–678, Mar. 2006.

[11] W. Zhang and B. D. Rao, "Combining independent component analysis with geometric information and its application to speech processing," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2009, pp. 3065–3068.

[12] L. Lu, H. J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 7, pp. 504–516, Oct. 2002.

[13] R. Huang and J. H. L. Hansen, "Advances in unsupervised audio classification and segmentation for the broadcast news and NGSW corpora," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 907–919, May 2006.

[14] H. L. V. Trees, *Optimum Array Processing*. New York: Wiley, 2002.

[15] S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 425–437, Sep. 1997.

[16] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[17] H. Li and T. Adalı, "A class of complex ICA algorithms based on the kurtosis cost function," *IEEE Trans. Neural Netw.*, vol. 19, no. 3, pp. 408–420, 2008.

[18] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Audio Process.*, vol. ASSP-32, no. 12, pp. 1109–1121, Dec. 1984.
[19] E. Bingham and A. Hyvärinen, "A fast fixed-point algorithm for independent component analysis of complex valued signals," *Int. J. Neural Syst.*, vol. 10, pp. 1–8, 2000.
[20] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 2, pp. 109–116, Mar. 2003.
[21] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 1997, vol. 2, pp. 1331–1334.
[22] A. I. Al-Shoshan, "Speech and music classification and separation: A review," *J. King Saud Univ.*, vol. 19, pp. 95–133, 2006.
[23] B. Sällberg, N. Grbić, and I. Claesson, "Complex-valued independent component analysis for online blind speech extraction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1624–1632, Nov. 2008.
[24] D. H. Brandwood, "A complex gradient operator and its application in adaptive array theory," in *Proc. IEE, Special Iss. Adaptive Arrays*, Feb. 1983, vol. 130, pp. 11–17.
[25] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York: Springer, 2006.
[26] M. M. Goulding and J. S. Bird, "Speech enhancement for mobile telephony," *IEEE Trans. Veh. Technol.*, vol. 39, no. 4, pp. 316–326, Nov.. 1990.
[27] [Online]. Available: http://www.utdallas.edu/research/utdrive/UT-Drive-Website.htm
[28] H. Abut, J. Hansen, and K. Takeda, *DSP for In-Vehicle and Mobile Systems*. New York: Springer, 2004.
[29] "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU, 2000, ITU-T Rec. 862.
[30] J. S. Garofolo, *Getting Started with the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database*. Gaithersburg, MD: NIST, 1988.
[31] G. Yan and H. Fan, "A Newton-like algorithm for complex variables with applications in blind equalization," *IEEE Trans. Signal Process.*, vol. 48, no. 2, pp. 553–556, Feb. 2000.

Tao Yu (S'08) received the B.S. degree both in electrical engineering and mathematics from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2004. He is currently pursuing the Ph.D. degree in the Center for Robust Speech System, University of Texas at Dallas, Richardson.

He was a Teaching Assistant in the University of Colorado at Boulder in 2006. His research interests are array signal processing, speech enhancement, and robust speech recognition.

John H. L. Hansen (S'81–M'82–SM'93–F'07) received the B.S.E.E. degree from the College of Engineering, Rutgers University, New Brunswick, NJ, in 1982 and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1983 and 1988, respectively.

He joined the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UTD), Richardson, in the fall of 2005, where he is Professor and Department Head of Electrical Engineering, and holds the Distinguished University Chair in Telecommunications Engineering. He also holds a joint appointment as Professor in the School of Behavioral and Brain Sciences (Speech and Hearing). At UTD, he established the Center for Robust Speech Systems (CRSS) which is part of the Human Language Technology Research Institute. Previously, he served as Department Chairman and Professor in the Department of Speech, Language and Hearing Sciences (SLHS) and Professor in the Department of Electrical and Computer Engineering at University of Colorado at Boulder (1998–2005), where he cofounded the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory (RSPL) and continues to direct research activities in CRSS at UTD. His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, robust speech recognition with emphasis on spoken document retrieval, and in-vehicle interactive systems for hands-free human–computer interaction. He has supervised 51 (22 Ph.D., 29 M.S./M.A.) thesis candidates. He is author/coauthor of 370 journal and conference papers and eight textbooks in the field of speech processing and language technology, coauthor of the textbook *Discrete-Time Processing of Speech Signals*, (IEEE Press, 2000), coeditor of *DSP for In-Vehicle and Mobile Systems* (Springer, 2004), *Advances for In-Vehicle and Mobile Systems: Challenges for International Standards* (Springer, 2006), *In-Vehicle Corpus and Signal Processing for Driver Behavior* (Springer, 2008), and lead author of the report "The Impact of Speech Under 'Stress' on Military Speech Technology," (NATO RTO-TR-10, 2000)

Prof. Hansen was named IEEE Fellow for contributions in "Robust Speech Recognition in Stress and Noise" in 2007 and is currently serving as Member of the IEEE Signal Processing Society Speech Technical Committee (2005–2008; 2010–2013; elected Chair-elect in 2010), and Educational Technical Committee (2005–2008; 2008–2010). Previously, he has served as Technical Advisor to U.S. Delegate for NATO (IST/TG-01), IEEE Signal Processing Society Distinguished Lecturer (2005/2006), Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1992–1999), Associate Editor for the IEEE SIGNAL PROCESSING LETTERS (1998–2000), and Editorial Board Member for the IEEE SIGNAL PROCESSING MAGAZINE (2001–2003). He has also served as a Guest Editor of the October 1994 special issue on Robust Speech Recognition for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He has served on the Speech Communications Technical Committee for the Acoustical Society of America (2000–2003), and is serving as a member of the International Speech Communications Association (ISCA) Advisory Council. In 2010, he was elected an ISCA Fellow for contributions in speech processing and recognition. He was recipient of the 2005 University of Colorado Teacher Recognition Award as voted on by the student body. . He also organized and served as General Chair for ICSLP/Interspeech-2002: International Conference on Spoken Language Processing, September 16–20, 2002, and served as Co-Organizer and Technical Program Chair for IEEE ICASSP-2010, Dallas, TX.