Modulation-Domain Multichannel Kalman Filtering for Speech Enhancement

Wei Xue, *Member, IEEE*, Alastair. H. Moore, *Member, IEEE*, Mike Brookes, *Member, IEEE*, and Patrick A. Naylor, *Senior Member, IEEE*

***Abstract*—Compared with single-channel speech enhancement methods, multichannel methods can utilize spatial information to design optimal filters. Although some filters adaptively consider second-order signal statistics, the temporal evolution of the speech spectrum is usually neglected. By using linear prediction (LP) to model the inter-frame temporal evolution of speech, single-channel Kalman filtering (KF) based methods have been developed for speech enhancement. In this paper, we derive a multichannel KF (MKF) that jointly uses both interchannel spatial correlation and interframe temporal correlation for speech enhancement. We perform LP in the modulation domain, and by incorporating the spatial information, derive an optimal MKF gain in the short-time Fourier transform domain. We show that the proposed MKF reduces to the conventional multichannel Wiener filter if the LP information is discarded. Furthermore, we show that, under an appropriate assumption, the MKF is equivalent to a concatenation of the minimum variance distortion response beamformer and a single-channel modulation-domain KF and therefore present an alternative implementation of the MKF. Experiments conducted on a public head-related impulse response database demonstrate the effectiveness of the proposed method.**

***Index Terms*—Speech enhancement, microphone arrays, Kalman filtering, modulation domain.**

I. Introduction

INTERFERENCE from environmental noise brings great challenges to speech processing systems in speech communication, hearing aids, and automatic speech recognition. Speech enhancement aims to suppress the environmental noise without distorting the target speech. If the target speech and noise arrive from different directions, a microphone array can beneficially be deployed to capture the spatial diversity of the acoustic environment. It is widely recognized that, by additionally exploiting spatial diversity, multichannel speech enhancement methods can achieve better performance than single-channel methods.

Manuscript received December 16, 2017; revised April 19, 2018 and June 1, 2018; accepted June 4, 2018. Date of publication June 8, 2018; date of current version June 22, 2018. This work was supported by the Engineering and Physical Sciences Research Council E-LOBES project EP/M026698/1. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Richard Christian Hendriks. *(Corresponding author: Wei Xue.)*

The authors are with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: w.xue@imperial.ac.uk; alastair.h.moore@imperial.ac.uk; mike.brookes@imperial.ac.uk; p.naylor@imperial.ac.uk).

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/TASLP.2018.2845665

The clean speech signal is redundantly expressed in the multichannel observations up to a steering vector or a relative transfer function (RTF), which depends on the direction of the target speaker and which defines the spatial correlation of the multiple observations of the clean signal. This redundancy makes the target signal spatially predictable [1] from clean signal observations of more than one channel, therefore, the signals of multiple microphones can jointly be used to estimate a single target signal.

Given knowledge of the steering vector, or RTF, traditional beamforming-based methods [2]–[9], which originate from narrowband methods for radar and sonar applications [10]–[12], can be developed to recover the desired signal from the noisy observations. Although fixed beamforming methods such as delay and sum (DS) [13] that depend only on the array geometry have the advantage of simplicity, adaptive beamforming techniques including minimum variance distortion response (MVDR) [3], [4], [7] and linear constrained minimum variance (LCMV) [8] have received more attention, since they can adjust to the spatio-temporal characteristics of speech and noise. Beamforming can be effective in reducing the directional noise but is not sufficient especially when the noise field is diffuse. To further reduce the noise signal, post-filtering [14]–[17] and the generalized sidelobe canceller (GSC) [18], [19] have been proposed. In these approaches, the beamforming is used as a preprocessor to obtain an enhanced signal, and the residual noise is further reduced by single-channel speech enhancement or multichannel adaptive noise cancellation (ANC) in the post-filtering stage. With post-filtering processing, various types of noise including diffuse noise, directional interference, and sensor noise can be reduced effectively. In addition, it is shown that, the GSC is equivalent to the MVDR beamformer by using the DS in the beamforming stage [19], [20]. Another category of multichannel speech enhancement algorithms is multichannel Wiener filtering (MWF) [21]–[23], which can operate without explicit knowledge of the steering vector or RTF, and estimates the target signal under the minimum mean squared error (MMSE) criterion. Speech distortion weighted MWF (SDW-MWF) [24] has also been developed to achieve a trade-off between noise reduction and speech distortion. It is shown in [25] that under a single-source assumption, the MWF can be expressed as a MVDR beamformer followed by a single-channel Wiener filter.

Since the speech signal can be expressed as an auto-regressive (AR) process, the signal in each frame is temporally correlated with the signals of previous frames. Therefore, the target signal

is not only spatially predictable from multichannel observations, but also temporally predictable from the previous frames. Although the temporal characteristics of speech or noise have been considered by some conventional methods such as MVDR, post-filtering and MWF to design adaptive filters, they are mainly used to compute the second order statistics (SOS) which inherently adopt a short-time stationary process assumption, and the dynamics of the speech signal over short-time frames are not considered.

In fact, by using linear prediction (LP) to model the temporal evolution of speech, many single-channel Kalman filtering (KF) based speech enhancement methods have been proposed. Since the pioneering work in [26], single-channel KF based methods have gradually developed from the time domain [27], [28] to the short-time Fourier transform (STFT) domain [29], [30] and modulation domain [31]–[39]. Since it has been found that there is almost no correlation in the successive phase samples [30], the temporal correlation is normally represented in the magnitude spectrum and using the amplitude spectrum for LP have received more attention. In addition, additional psychoacoustic and physiological advantages of modulation-domain speech processing [40], [41] can be further exploited to improve the speech enhancement performance. Therefore, the modulation-domain methods have become mainstream.

In modulation-domain processing, the time-varying envelope amplitude in each frequency bin is regarded as a time-domain signal in its own right. In the single-channel modulation-domain Kalman filtering (MDKF) methods [31]–[39], a modulation-domain state vector is defined to represent the amplitude estimation of the clean speech. Based an LP model of clean speech, the state vector of the current frame is firstly predicted, and then updated by combining with the modulation-domain noisy observation. A KF gain is computed based on the MMSE criterion to combine the modulation-domain LP estimation with the noisy observation.

An early approach involving the KF for multichannel speech enhancement was presented in Chapter 5.9 of [42], where a time-domain approach was derived. However, as is discussed therein, the approach requires blind estimation of the room impulse responses (RIRs) in noisy conditions, which remains an open problem, and so the practicality of the approach is limited. As a result, the method was only introduced conceptually without experimental evaluation. It is relevant to note that the KF has been used in other areas of multichannel speech processing such as speech dereverberation [43], [44] and speaker tracking [45], [46].

In this paper, we develop a modulation-domain KF for the multichannel case, and propose a novel multichannel Kalman filtering (MKF) algorithm for speech enhancement. The proposed MKF operates in both the modulation domain and the STFT domain, so that it can jointly utilize the inter-frame temporal correlation and the inter-channel spatial correlation to estimate the target clean speech. As in the classical KF [47], the key idea of the MKF is to find an optimal MKF gain that uses the LP estimation derived from the dynamic model and previous estimates in a weighted combination with the estimation derived from the measurement model and observations. It is expected that higher weight is given to the LP estimation in noise-dominated time-frequency (TF) bins, and to the observation in speech-dominated TF bins.

The LP estimation is obtained, by first computing a modulation-domain prediction from the previous estimates, and then transforming this into the STFT-domain. To exploit the spatial information, we incorporate the multichannel noisy observations and calculate the optimal MKF gain in the STFT domain according to an MMSE criterion. We report experiments using a head-related impulse response (HRIR) database [48], and the results show that the proposed MKF outperforms conventional methods in a range of noisy and reverberant environments. We show that the proposed MKF reduces to the conventional multichannel Wiener filter (MWF) if the LP information is discarded and also show that, under appropriate conditions, the MKF is equivalent to a MVDR beamformer followed by a single-channel MDKF.

This paper is an extension of our previous work in [49]. The remainder of the paper is organized as follows. In Section II, we introduce the signal model and some required assumptions. In Section III, the proposed method is described, including the MKF model and the derivation of the MKF. We also discuss the relationship between the MKF and the MWF. Then in Section V, we show that, by reformulating the proposed MKF, it can be expressed as the concatenation of an MVDR and a single-channel MDKF post-filter, such that the optimality of the concatenation is shown, and an alternative implementation of the MKF is presented. Finally in Section VI, we conduct experiments in different noisy and reverberant conditions, and show the effectiveness of the proposed method.

Table I in the next page summarizes the notations used for vectors and matrices in this paper.

II. Signal Model and Assumptions

We consider a noisy and reverberant environment which includes a target talker and an M-element microphone array. The target speech of each microphone is contaminated by diffuse acoustic noise, sensor noise, and potentially also disturbing signals from interfering sources. We neglect the details of different noise components and treat them as a combination, so that, in the complex-valued STFT domain, the noisy signal vector $\mathbf{y}(n,k)$ captured by the microphone array in the n-th frame and k-th frequency can be written as

$$\mathbf{y}(n,k) = \mathbf{x}(n,k) + \mathbf{v}(n,k), \tag{1}$$

where $\mathbf{y}(n,k) = [Y_1(n,k),\, Y_2(n,k),\, \ldots,\, Y_M(n,k)]^T$, and $Y_m(n,k)$ for $m = 1, 2, ..., M$ is the STFT of the noisy signal of the m-th microphone. $\mathbf{x}(n,k)$ and $\mathbf{v}(n,k)$ are the target speech and additive noise vectors, respectively, and have the same form as $\mathbf{y}(n,k)$. We assume that the speech and noise signals are uncorrelated.

By taking the speech signal of the first microphone $X_1(n,k)$ as reference, $\mathbf{y}(n,k)$ can also be expressed as

$$\mathbf{y}(n,k) = \mathbf{d}(k)X_1(n,k) + \mathbf{v}(n,k), \tag{2}$$

TABLE I
NOTATION FOR VECTORS AND MATRICES

$\mathbf{y}(n,k), \mathbf{x}(n,k), \mathbf{v}(n,k)$	Multichannel STFT-domain signal vectors of observation, speech and noise, respectively.
$\mathbf{d}(k)$	RTF vector.
$\mathbf{x}_1(n,k)$	State vector of MKF, signal vector of the reference channel.
$\mathbf{B}(k)$	State transition matrix.
$\mathbf{u}$	Column vector $[1,\ 0,\ \ldots,\ 0]^T$ with dimension $P \times 1$.
$\mathbf{z}(n,k)$	STFT-domain vector of the pre-processed signal after MWF noise reduction.
$\mathbf{Q}(k)$	$M \times P$ measurement matrix.
$\hat{\mathbf{x}}_1(n\|n-1,k)$, $\hat{\mathbf{x}}_1(n\|n,k)$	STFT-domain LP estimation and MMSE estimation of the state vector in the current frame, respectively.
$\hat{\mathbf{a}}_1(n\|n-1,k)$, $\hat{\mathbf{a}}_1(n\|n,k)$	Modulation-domain LP estimation and MMSE estimation of the state vector amplitude in the current frame, respectively.
$\hat{\mathbf{\Phi}}(n,k)$	Diagonal phase matrix containing the complex exponential phase of $\mathbf{z}(n,k)$.
$\mathbf{e}(n\|n-1,k)$, $\mathbf{e}(n\|n,k)$	STFT-domain LP and MMSE estimation error vectors in the current frame, respectively.
$\mathbf{R}_{ee}(n\|n-1,k)$, $\mathbf{R}_{ee}(n\|n,k)$	Covariance matrices of STFT-domain LP and MMSE estimation error vectors, respectively.
$\mathbf{G}(n,k)$, $\mathbf{K}(n,k)$, $\tilde{\mathbf{K}}(n,k)$	MKF gain, MDKF gain, and an STFT-domain KF gain defined in (46), respectively.
$\mathbf{R}_{xx}(n,k)$, $\mathbf{R}_{vv}(n,k)$	Speech covariance matrix and noise covariance matrix, respectively.
$\boldsymbol{\eta}(n\|n-1,k)$, $\boldsymbol{\eta}(n\|n,k)$	Modulation-domain LP and MMSE estimation error vectors in the current frame, respectively.
$\mathbf{R}_{\eta\eta}(n\|n-1,k)$, $\mathbf{R}_{\eta\eta}(n\|n,k)$	Covariance matrices of modulation-domain LP and MMSE estimation error vectors, respectively.
$\mathbf{h}_{\mathrm{mvdr}}(n,k)$	MVDR beamformer.

where

$$\mathbf{d}(k) = [d_1(n,k),\ d_2(n,k),\ \ldots,\ d_M(n,k)]^T \tag{3}$$

is a RTF vector, and $d_m(n,k)$ for $m = 1, 2, ..., M$ is the RTF between the m-th and first channel. Without loosing generality, we take the first channel as reference, $d_1(n,k) = 1$. As is common in the beamforming and post-filtering literature, we assume that the RTF is known a priori, or has already been estimated (by e.g. [50]).

The goal of the multichannel speech enhancement problem is to estimate the clean reverberant signal $X_1(n,k)$ based on the multichannel noisy observations $\mathbf{y}(n,k)$.

III. MKF STATE-SPACE MODEL

A new MKF algorithm which estimates the clean signal by jointly using the temporal correlation of speech and spatial correlation of multichannel signals is proposed. The MKF follows a similar structure with the single-channel MKDF [31]–[39], which first computes the LP estimate based on the temporal evolution model of speech, and obtains the final estimation by incorporating with the noisy observation. The derivations in the multichannel case will be presented in detail. We begin with introducing the MKF state-space model in this section, and then, in Section IV, derive the MKF to estimate the hidden state which represents the target speech.

As in the general KF [47], the state-space model of the proposed MKF includes a LP equation which describes the temporal evolution of the clean speech, and a measurement equation which describes the relationship between the clean speech and observation.

A. LP Equation

To model the temporal evolution of speech, we temporarily ignore the multichannel spatial information, and only consider the clean signal of the first channel. As the amplitude spectrum shows much larger temporal correlation over adjacent frames than the phase spectrum [51], due to the psychoacoustic and physiological advantages of modulation-domain speech processing [40], [41], the LP model is formulated in the modulation domain.

In the modulation-domain, by modelling the temporal evolution of speech as an AR model, the relationship between the speech signal in the current frame and previous frames is expressed as

$$A_1(n,k) = -\sum_{p=1}^{P} b_{p,k} A_1(n-p,k) + W(n,k), \tag{4}$$

where $A_1(n,k)$ indicates the magnitude of the STFT-domain signal $X_1(n,k)$, and $b_{p,k}$ for $p = 1, 2, ..., P$ are the LP coefficients [52] of the modulation-domain signal in the k-th frequency bin. $W(n,k)$ is a Gaussian distributed random excitation signal with variance δ_W^2.

Similar to [31], we define the state vector of MKF to be the signal vector of the reference channel given by $\mathbf{x}_1(n,k) = [X_1(n,k),\ X_1(n-1,k),\ \ldots,\ X_1(n-P+1,k)]^T$, then the LP equation of the MKF is

$$\mathbf{a}_1(n,k) = \mathbf{B}(k)\mathbf{a}_1(n-1,k) + \mathbf{u}W(n,k), \tag{5}$$

where $\mathbf{a}_1(n,k) = [A_1(n,k),\ A_1(n-1,k),\ \ldots,\ A_1(n-P+1,k)]^T$ is a vector containing the amplitude of each element of the state vector $\mathbf{x}_1(n,k)$.

In (5), $\mathbf{B}(k)$ is the $P \times P$ state transition matrix defined as

$$\mathbf{B}(k) = \begin{bmatrix} -b_{1,k} & -b_{2,k} & \ldots & -b_{p-1,k} & -b_{p,k} \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix}, \tag{6}$$

and $\mathbf{u} = [1,\ 0,\ \ldots,\ 0]^T$.

In practice, $\mathbf{B}(k)$ and δ_W^2 are unknown but can be estimated via LP analysis in the modulation frames [31], [37] by using the speech signal after noise reduction preprocessing $\mathbf{z}(n,k)$. In each frequency bin, the temporal sequence of spectral amplitudes, $\mathbf{z}(n,k)$, is segmented into overlapping "modulation frames" [37]. Here $\mathbf{z}(n,k)$ is defined in the STFT-domain as $\mathbf{z}(n,k) = [Z_1(n,k), Z_1(n-1,k), \ldots, Z_1(n-P+1,k)]^T$, where $Z_1(n,k)$ is the preprocessed signal of the first channel. For single-channel MDKF methods, $\mathbf{z}(n,k)$ is usually computed by single-channel enhancers such as [53],

[54]. In the multichannel case, since multiple microphones are available, we take $\mathbf{z}(n,k)$ as the output of the MWF, which is realized as a MVDR beamformer followed by a single-channel Wiener post-filter [55]. Details of the implementation of the preprocessing are given in Section VI-A, and the MVDR beamformer weights are given in (29).

By incorporating the LP information of speech for speech enhancement, the MKF implicitly assumes the dependency between STFT frames arises only from the target speech signal. However, correlation between frames in the STFT and modulation domain may be introduced by the noise, especially when the frame length of STFT is short and frames are overlapped as discussed in [32]. The inter-frame correlation and the temporal evolution of the noise can be modelled explicitly by incorporating the noise into the KF state as has been done for single-channel KF based methods [31]–[33], [36]. Our current implementation does not include the noise signal in the KF state but nevertheless, as will be shown in the experiments, it can generally perform better than conventional methods.

B. Measurement Equation

The multichannel spatial information which is not exploited in the LP equation is considered now in order to define the measurement equation. Since the spatial information is carried primarily by the phase spectrum, according to the signal model in (2), the measurement equation is defined in the STFT domain, as

$$\begin{aligned}\mathbf{y}(n,k) &= \mathbf{d}(k)X_1(n,k) + \mathbf{v}(n,k)\\ &= \mathbf{d}(k)\mathbf{u}^T\mathbf{x}_1(n,k) + \mathbf{v}(n,k)\\ &= \mathbf{Q}(k)\mathbf{x}_1(n,k) + \mathbf{v}(n,k),\end{aligned} \tag{7}$$

where $\mathbf{Q}(k) = \mathbf{d}(k)\mathbf{u}^T$ is an $M \times P$ measurement matrix. Since $\mathbf{Q}(k)$ is a function of $\mathbf{d}(k)$, unlike single-channel MKDF methods such as [31], the RTF which consists of the spatial information is integrated into the MKF model.

For simplicity of representation, in the rest of this paper, the frequency index "k" will be omitted. We would like to note that the RTF $\mathbf{d}$ and measurement matrix $\mathbf{Q}$ are frequency-dependent, and $\mathbf{u}$ is a constant vector.

IV. Derivation of MKF

We begin by noting that the LP equation and measurement equation are defined in the modulation domain and STFT domain respectively, and that taking the magnitude from the STFT spectrum is non-linear. Consequently, the state vector cannot be estimated using the conventional linear KF framework. In this section, based on the state-space model of MKF, an MKF is derived to estimate the state vector that represents the target clean signal.

The framework of the proposed MKF is illustrated in Fig. 1. As with the conventional KF, the proposed MKF comprises an LP step and an update step. In the LP step, an *a priori* modulation-domain LP estimate $\hat{\mathbf{a}}_1(n|n-1)$ is first calculated, and then transformed into the STFT domain $\hat{\mathbf{x}}_1(n|n-1)$. In the update step, by incorporating the noisy observation $\mathbf{y}(n)$, we derive an optimal MKF gain, and compute the *a posteriori* state vector estimate $\hat{\mathbf{x}}_1(n|n)$. The clean signal in the reference channel is estimated by taking the first element of the state vector, and transforming it into the time domain using the inverse STFT.

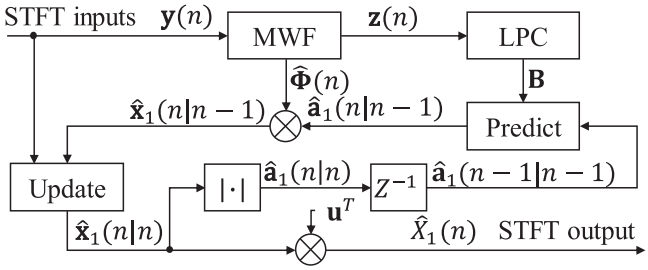

Fig. 1. MKF framework. $\hat{\mathbf{\Phi}}(n)$ is a diagonal phase matrix containing the complex exponential phase of $\mathbf{z}(n)$.

In the following, the details of the proposed MKF will be described.

A. STFT-domain State Vector Prediction

The dynamic model of speech in (5) describes the temporal relationship between the state vector in the current frame and the state vector in the previous frame. Similar to [31], given the MMSE estimate of the state vector of the previous frame, $\hat{\mathbf{x}}_1(n-1|n-1)$, the amplitude of the state vector in the current frame can be predicted as

$$\hat{\mathbf{a}}_1(n|n-1) = \mathbf{B}\hat{\mathbf{a}}_1(n-1|n-1), \tag{8}$$

where $\hat{\mathbf{a}}_1(n-1|n-1)$ is a vector containing the amplitude of each element of the MMSE estimate $\hat{\mathbf{x}}_1(n-1|n-1)$.

The $\hat{\mathbf{a}}_1(n|n-1)$ in (8) is defined in the modulation domain. To obtain the STFT-domain LP estimate $\hat{\mathbf{x}}_1(n|n-1)$, we further insert the phase of $\mathbf{z}(n)$ into $\hat{\mathbf{a}}_1(n|n-1)$, which is the approximation of the phase of clean speech, such that

$$\hat{\mathbf{x}}_1(n|n-1) = \hat{\mathbf{\Phi}}(n)\hat{\mathbf{a}}_1(n|n-1), \tag{9}$$

where $\mathbf{\Phi}(n)$ is a $P \times P$ diagonal matrix whose diagonal elements are the complex-valued exponential phase of $\mathbf{x}_1(n)$, and $\hat{\mathbf{\Phi}}(n)$ is computed based on the MWF output $\mathbf{z}(n)$, as

$$\hat{\mathbf{\Phi}}_{i,i}(n) = \frac{Z_1(n-i+1)}{|Z_1(n-i+1)|}. \tag{10}$$

We define the STFT-domain LP estimation error vector $\mathbf{e}(n|n-1)$ between $\mathbf{x}_1(n)$ and $\hat{\mathbf{x}}_1(n|n-1)$, which will be required in (14) below, as

$$\mathbf{e}(n|n-1) = \hat{\mathbf{x}}_1(n|n-1) - \mathbf{x}_1(n). \tag{11}$$

The prediction in (8) can, in rare circumstances, result in a negative amplitude for $\hat{\mathbf{a}}_1(n|n-1)$ which will, in turn, add a phase shift of 180° in (9). Negative values can be prevented by following the prediction by a half-wave rectifier, $\max\{\hat{\mathbf{a}}_1(n|n-1), \mathbf{0}\}$. However, since the half-wave rectifier is non-linear, in order to avoid complicating the theoretical analysis, we have not done this in the current work. In our experiments, we have

found that negative values of $\hat{\mathbf{a}}_1(n|n-1)$ are rare and occur in $<0.5\%$ of the time-frequency (T-F) cells, generally cells that are dominated by noise.

B. STFT-domain State Vector Update

Although the clean speech can be estimated by LP based on the dynamic model, this does not exploit the spatial information in the multichannel signal. In addition, the dynamic model incorporates a prediction error which, in some frames, may be large. The LP estimation can be improved by incorporating the new multichannel observations which, although they might be noisy, provide the new instantaneous and redundant information about the clean speech signals in different microphones.

For each new frame, by combining the estimations from STFT-domain LP estimation $\hat{\mathbf{x}}_1(n|n-1)$ in (9) and the multichannel noisy observation $\mathbf{y}(n)$, we update the MKF state vector by:

$$\hat{\mathbf{x}}_1(n|n) = \hat{\mathbf{x}}_1(n|n-1) + \mathbf{G}(n)[\mathbf{y}(n) - \mathbf{Q}\hat{\mathbf{x}}_1(n|n-1)], \tag{12}$$

where $\mathbf{G}(n)$ is the $P \times M$ MKF gain.

The error vector $\mathbf{e}(n|n)$ between $\mathbf{x}_1(n)$ and the new estimate $\hat{\mathbf{x}}_1(n|n)$ is computed as

$$\mathbf{e}(n|n) = \hat{\mathbf{x}}_1(n|n) - \mathbf{x}_1(n). \tag{13}$$

Combining (11) and (12), $\mathbf{e}(n|n)$ becomes

$$\begin{aligned}
&\mathbf{e}(n|n)\\
&= \hat{\mathbf{x}}_1(n|n-1) - \mathbf{x}_1(n) + \mathbf{G}(n)[\mathbf{y}(n) - \mathbf{Q}\hat{\mathbf{x}}_1(n|n-1)]\\
&= \mathbf{e}(n|n-1) + \mathbf{G}(n)[\mathbf{v}(n) - \mathbf{Q}\mathbf{e}(n|n-1)]\\
&= [\mathbf{I} - \mathbf{G}(n)\mathbf{Q}]\mathbf{e}(n|n-1) + \mathbf{G}(n)\mathbf{v}(n).
\end{aligned} \tag{14}$$

To compute the MKF gain, we first define a cost function under the MMSE criterion as

$$J[\mathbf{G}(n)] = \mathrm{tr}[\mathbf{R}_{ee}(n|n)], \tag{15}$$

where $\mathbf{R}_{ee}(n|n) = \mathbb{E}\{\mathbf{e}(n|n)\mathbf{e}^H(n|n)\}$ and $\mathbb{E}\{\cdot\}$ is the expectation operator. The optimal MKF gain $\mathbf{G}(n)$ is found by minimizing $J[\mathbf{G}(n)]$.

Since $\mathbf{e}(n|n-1)$ is estimated from $\mathbf{v}(n-1)$ and $\mathbf{x}(n|n-1)$, it is uncorrelated with the current noise signal $\mathbf{v}(n)$. By multiplying (14) by its conjugate transpose and taking expectations, we therefore have

$$\begin{aligned}
\mathbf{R}_{ee}(n|n) = {}& [\mathbf{I} - \mathbf{G}(n)\mathbf{Q}]\mathbf{R}_{ee}(n|n-1)[\mathbf{I} - \mathbf{G}(n)\mathbf{Q}]^H\\
&+ \mathbf{G}(n)\mathbf{R}_{vv}(n)\mathbf{G}^H(n),
\end{aligned} \tag{16}$$

where $\mathbf{R}_{ee}(n|n-1) = \mathbb{E}\{\mathbf{e}(n|n-1)\mathbf{e}^H(n|n-1)\}$ is the STFT-domain LP estimation error covariance matrix, and $\mathbf{R}_{vv}(n) = \mathbb{E}\{\mathbf{v}(n)\mathbf{v}^H(n)\}$ is the noise covariance matrix, which can be estimated by several existing methods, such as [56]–[58].

From [59], the derivative of $J[\mathbf{G}(n)]$ over $\mathbf{G}(n)$ can be computed as

$$\begin{aligned}
&\frac{\partial J[\mathbf{G}(n)]}{\partial \mathbf{G}(n)}\\
&= -2[\mathbf{I} - \mathbf{G}(n)\mathbf{Q}]\mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H + 2\mathbf{G}(n)\mathbf{R}_{vv}(n).
\end{aligned} \tag{17}$$

By setting the derivative to zero, we obtain the optimal MKF gain:

$$\mathbf{G}(n) = \mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H[\mathbf{Q}\mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H + \mathbf{R}_{vv}(n)]^{-1}. \tag{18}$$

According to (18), besides the measurement matrix $\mathbf{Q}$ and the estimated $\mathbf{R}_{vv}(n)$, the optimal MKF gain is a function of the STFT-domain LP estimation error covariance matrix $\mathbf{R}_{ee}(n|n-1)$, which is unknown. Estimating $\mathbf{R}_{ee}(n|n-1)$ will be described in Section IV-C. We note that if the noise level is low, the condition number of $[\mathbf{Q}\mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H + \mathbf{R}_{vv}(n)]$ will be large, since $\mathbf{Q}\mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H$ is actually of rank one from the definition of $\mathbf{Q}$ in (7). In this case, the matrix inverse in (18) can be replaced by its pseudo-inverse to avoid inverting a near-singular matrix.

From (12) and (18) we can see that the MKF gain adjusts the weighting between the LP estimation and the estimation from the observations according to the noise level, which is the same with the classical KF [47]. When the noise level is high, large values of the $\mathbf{R}_{vv}(n)$ elements reduce the values of the $\mathbf{G}(n)$ elements, and consequently the updated state vector more heavily favours the LP estimation. In the opposite case, since the observations provide a more accurate description of the clean speech, the effect of new observations is increased in the updated state vector.

C. Estimating $\mathbf{R}_{ee}(n|n-1)$

The matrix $\mathbf{R}_{ee}(n|n-1)$ in (18) is unknown and will be estimated in this subsection. A key feature of beamforming is that it can estimate not only the amplitude but also the phase of the clean speech signal given the RTF. Here, we use the phase of $\hat{\mathbf{x}}_1(n|n-1)$, which is same as that of the MVDR output, to approximate the phase of the clean signal $\mathbf{x}_1(n)$. Similar to (9), we can then rewrite $\mathbf{x}_1(n)$ as

$$\mathbf{x}_1(n) = \mathbf{\Phi}(n)|\mathbf{x}_1(n)|. \tag{19}$$

Assuming $\hat{\mathbf{\Phi}}(n)$ is a good approximation of $\mathbf{\Phi}(n)$, we have $\mathbf{\Phi}(n) = \hat{\mathbf{\Phi}}(n)$. From (11), obtain

$$\begin{aligned}
\mathbf{e}(n|n-1) &= \hat{\mathbf{\Phi}}(n)[|\mathbf{x}_1(n)| - \hat{\mathbf{a}}_1(n|n-1)]\\
&= \hat{\mathbf{\Phi}}(n)\boldsymbol{\eta}(n|n-1),
\end{aligned} \tag{20}$$

where $\boldsymbol{\eta}(n|n-1) = |\mathbf{x}_1(n)| - \hat{\mathbf{a}}_1(n|n-1)$ is the modulation-domain LP estimation error vector.

We further define $\mathbf{R}_{\eta\eta}(n|n-1) = \mathbb{E}\{\boldsymbol{\eta}(n|n-1)\boldsymbol{\eta}^H(n|n-1)\}$ as the modulation-domain LP estimation error covariance matrix. Since $\boldsymbol{\eta}(n|n-1)$ is defined in the single-channel modulation domain, the covariance matrix can be updated as in the

Algorithm 1: MKF.

$\hat{\mathbf{x}}_1(n|n) = [Y_1(n) \ \ldots \ Y_1(n-P+1)]^H, n \leq P.$
for $n = P+1$ *to* N **do**
- a) Determine $\hat{\mathbf{a}}_1(n|n-1)$ and $\mathbf{R}_{\eta\eta}(n|n-1)$ from (8) and (21);
- b) Determine $\hat{\mathbf{x}}_1(n|n-1)$ and $\mathbf{R}_{ee}(n|n-1)$ from (9) and (22);
- c) Compute the MKF gain $\mathbf{G}(n)$ from (18);
- d) Update the state vector $\hat{\mathbf{x}}_1(n|n)$ using (12);
- e) Update $\mathbf{R}_{ee}(n|n)$ and $\mathbf{R}_{\eta\eta}(n|n)$ from (25) and (26);

end
N is the number of frames.

conventional MDKF [31], as

$$\mathbf{R}_{\eta\eta}(n|n-1) = \mathbf{B}\mathbf{R}_{\eta\eta}(n-1|n-1)\mathbf{B}^H + \delta_W^2\,\mathbf{u}\mathbf{u}^H. \quad (21)$$

From (20), by integrating the phase information, the STFT-domain LP error covariance matrix $\mathbf{R}_{ee}(n|n-1)$ is updated using:

$$\mathbf{R}_{ee}(n|n-1) = \hat{\mathbf{\Phi}}(n)\mathbf{R}_{\eta\eta}(n|n-1)\hat{\mathbf{\Phi}}^H(n). \quad (22)$$

The $\mathbf{R}_{ee}(n|n-1)$ is substituted into (18) to compute the optimal MKF gain $\mathbf{G}(n)$. Once the state vector in (12) is computed given $\mathbf{G}(n)$, $\mathbf{R}_{ee}(n|n)$ in (16) can be updated. We first rewrite (16) as

$$\begin{aligned}\mathbf{R}_{ee}(n|n) &= [\mathbf{I} - \mathbf{G}(n)\mathbf{Q}]\mathbf{R}_{ee}(n|n-1) \\ &\quad - [\mathbf{I} - \mathbf{G}(n)\mathbf{Q}]\mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H\mathbf{G}^H(n) \\ &\quad + \mathbf{G}(n)\mathbf{R}_{vv}(n)\mathbf{G}^H(n). \end{aligned} \quad (23)$$

Since the derivative in (17) is zero, the second term on the right-hand side of (23) can be written as

$$[\mathbf{I} - \mathbf{G}(n)\mathbf{Q}]\mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H\mathbf{G}^H(n) = \mathbf{G}(n)\mathbf{R}_{vv}(n)\mathbf{G}^H(n). \quad (24)$$

It follows that,

$$\mathbf{R}_{ee}(n|n) = [\mathbf{I} - \mathbf{G}(n)\mathbf{Q}]\mathbf{R}_{ee}(n|n-1). \quad (25)$$

Noting that $\hat{\mathbf{\Phi}}^{-1}(n) = \hat{\mathbf{\Phi}}^H(n)$,

$$\mathbf{R}_{\eta\eta}(n|n) = \hat{\mathbf{\Phi}}^H(n)\mathbf{R}_{ee}(n|n)\hat{\mathbf{\Phi}}(n). \quad (26)$$

The steps of the proposed MKF for each frequency k are summarized in Algorithm 1. The clean signal estimation of the first channel $\hat{X}_1(n)$ is as $\mathbf{u}^T\hat{\mathbf{x}}_1(n|n)$.

D. Relationship to the MWF

We now discuss the relationship between the proposed MKF and the MWF [21]. If the estimates from LP are not used, by setting $\hat{\mathbf{x}}_1(n|n-1) = 0$, $\mathbf{e}(n)$ in (11) becomes $-\mathbf{x}_1(n)$, and $\mathbf{Q}\mathbf{e}(n)$ becomes $-\mathbf{x}(n)$. As a result, in (18), $\mathbf{Q}\mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H$ becomes the speech covariance matrix $\mathbf{R}_{xx}(n) = \mathbb{E}\{\mathbf{x}(n)\mathbf{x}^H(n)\}$, and the MKF gain $\mathbf{G}(n)$ can be written as

$$\mathbf{G}(n) = \mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H[\mathbf{R}_{xx}(n) + \mathbf{R}_{vv}(n)]^{-1}. \quad (27)$$

Note that $\mathbf{u}^T\mathbf{d} = 1$, and $\hat{\mathbf{x}}_1(n)$ in (12) equals $\mathbf{G}(n)\mathbf{y}(n)$ since $\hat{\mathbf{x}}_1(n|n-1) = 0$, then the clean signal estimation in the reference channel can be expressed as

$$\begin{aligned}\hat{X}_1(n) &= \mathbf{u}^T\hat{\mathbf{x}}_1(n) \\ &= \mathbf{u}^T\mathbf{d}\mathbf{u}^T\hat{\mathbf{x}}_1(n) \\ &= \mathbf{u}^T\mathbf{Q}\mathbf{G}(n)\mathbf{y}(n) \\ &= \mathbf{u}^T\mathbf{R}_{xx}(n)[\mathbf{R}_{xx}(n) + \mathbf{R}_{vv}(n)]^{-1}\mathbf{y}(n) \\ &= \mathbf{h}_{\mathrm{mwf}}^H\mathbf{y}(n), \end{aligned} \quad (28)$$

where $\mathbf{h}_{\mathrm{mwf}} = [\mathbf{R}_{xx}(n) + \mathbf{R}_{vv}(n)]^{-1}\mathbf{R}_{xx}(n)\mathbf{u}$ is the MWF [3]. Therefore the MKF reduces to the MWF, and the proposed MKF can be seen as integrating the temporal evolution of speech into the conventional MWF.

V. Factorization of MKF and Alternative Implementation

In this section, we present a theoretical analysis of the MKF. We deduce that if we use the phase of the MVDR output to approximate the clean phase, the MKF, which is optimally derived under the MMSE criterion, can be theoretically expressed as a MVDR-MDKF. The MVDR-MDKF is a concatenation of an MVDR beamformer and a single-channel MDKF. It has been shown that the MVDR beamformer is the sufficient statistic of the clean signal using the multichannel observations [60], therefore, the MVDR beamformer is chosen to produce the input signal for MDKF post-filtering. Based on the analysis, an alternative implementation of the MKF is proposed.

A. MVDR-MDKF

In this subsection, we derive the output of the MVDR-MDKF.

1) MVDR Beamforming: Given the RTF, $\mathbf{d}$, and the noise covariance matrix, $\mathbf{R}_{vv}(n)$, the MVDR beamformer is designed as [61]:

$$\mathbf{h}_{\mathrm{mvdr}}(n) = \frac{\mathbf{R}_{vv}^{-1}(n)\mathbf{d}}{\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbf{d}}. \quad (29)$$

Given the multichannel noisy signal $\mathbf{y}(n)$, according to the signal model (2), the MVDR output $\mathcal{Z}_1(n)$ is obtained as

$$\begin{aligned}\mathcal{Z}_1(n) &= \mathbf{h}_{\mathrm{mvdr}}^H(n)\mathbf{y}(n) \\ &= \mathbf{h}_{\mathrm{mvdr}}^H(n)\mathbf{x}(n) + \mathbf{h}_{\mathrm{mvdr}}^H(n)\mathbf{v}(n) \\ &= \frac{\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)}{\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbf{d}}\mathbf{d}X_1(n) + V_o(n) \\ &= X_1(n) + V_o(n), \end{aligned} \quad (30)$$

where $V_o(n) = \mathbf{h}_{\mathrm{mvdr}}^H(n)\mathbf{v}(n)$ is the residual noise at the MVDR output.

2) Single-channel MDKF Post-Filtering: In the post-filtering step, the single-channel MDKF is applied to the MVDR output $\mathcal{Z}_1(n)$. According to [31], for $\mathcal{Z}_1(n)$, the state-space

model of the single-channel MDKF is written as

$$|\mathbf{x}_1(n)| = \mathbf{B}|\mathbf{x}_1(n-1)| + \mathbf{u}W(n) \tag{31}$$

$$|\mathcal{Z}_1(n)| = \mathbf{u}^T|\mathbf{x}_1(n)| + |V_o(n)|, \tag{32}$$

and the state vector can be estimated iteratively by

$$\mathbf{R}_{\eta\eta}(n|n-1) = \mathbf{B}\mathbf{R}_{\eta\eta}(n-1|n-1)\mathbf{B}^H + \delta_W^2\,\mathbf{u}\mathbf{u}^T \tag{33}$$

$$\hat{\mathbf{a}}_1(n|n-1) = \mathbf{B}\hat{\mathbf{a}}_1(n-1|n-1) \tag{34}$$

$$\mathbf{K}(n) = \mathbf{R}_{\eta\eta}(n|n-1)\mathbf{u} \times [\delta_{V_o}^2 + \mathbf{u}^T\mathbf{R}_{\eta\eta}(n|n-1)\mathbf{u}]^{-1} \tag{35}$$

$$\mathbf{R}_{\eta\eta}(n|n) = [\mathbf{I} - \mathbf{K}(n)\mathbf{u}^T]\mathbf{R}_{\eta\eta}(n|n-1) \tag{36}$$

$$\hat{\mathbf{a}}_1(n|n) = \hat{\mathbf{a}}_1(n|n-1) + \mathbf{K}(n)[|\mathcal{Z}_1(n)| - \mathbf{u}^T\hat{\mathbf{a}}_1(n|n-1)]. \tag{37}$$

The clean signal estimate of the first channel $\hat{X}_1(n)$ is obtained by inserting the phase of $\mathcal{Z}_1(n)$ into $\mathbf{u}^T\hat{\mathbf{a}}_1(n|n)$. The phase of $\mathcal{Z}_1(n)$ is equal to the phase of the MWF output $Z_1(n)$ in (10), since the MWF is implemented as a MVDR and a single-channel Wiener post-filter. Therefore,

$$\hat{X}_1(n) = \hat{\mathbf{\Phi}}_{1,1}(n)\mathbf{u}^T\hat{\mathbf{a}}_1(n|n) = \mathbf{u}^T\hat{\mathbf{\Phi}}(n)\hat{\mathbf{a}}_1(n|n), \tag{38}$$

where $\hat{\mathbf{\Phi}}_{1,1}(n)$ is the upper left element of $\hat{\mathbf{\Phi}}(n)$.

B. Factorization of MKF Gain

According to the definition of $\mathbf{Q}$ in (7), we can simplify $\mathbf{Q}\mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H$ as

$$\mathbf{Q}\mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H = \mathbf{d}\mathbf{u}^T\mathbf{R}_{ee}(n|n-1)\mathbf{u}\mathbf{d}^H = \delta_e^2(n|n-1)\mathbf{d}\mathbf{d}^H, \tag{39}$$

where

$$\delta_e^2(n|n-1) = \mathbf{u}^T\mathbf{R}_{ee}(n|n-1)\mathbf{u} \tag{40}$$

is a scalar representing the first element of $\mathbf{R}_{ee}(n|n-1)$.

By applying the Sherman-Morrison-Woodbury formula [62]

$$(\mathbf{F}^{-1} + \mathbf{F}\mathbf{u}^{-1}\mathbf{F}^H)^{-1} = \mathbf{F} - \mathbf{A}\mathbf{F}(\mathbf{u} + \mathbf{F}^H\mathbf{A}\mathbf{F})^{-1}\mathbf{F}^H\mathbf{F}, \tag{41}$$

and letting $\mathbf{A} = \mathbf{R}_{vv}^{-1}(n)$, $\mathbf{F} = \delta_e(n|n-1)\mathbf{d}$, $\mathbf{u} = 1$, the MKF gain in (18) can be factorized as

$$\begin{aligned}\mathbf{G}(n) &= \mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H \times \left[\mathbf{R}_{vv}^{-1}(n) - \frac{\delta_e^2(n|n-1)\mathbf{R}_{vv}^{-1}(n)\mathbf{d}\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)}{1+\delta_e^2(n|n-1)\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbf{d}}\right] \\ &= \mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H\mathbf{R}_{vv}^{-1}(n)\left[1 - \frac{\delta_e^2(n|n-1)\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbf{d}}{1+\delta_e^2(n|n-1)\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbf{d}}\right] \\ &= \frac{\mathbf{R}_{ee}(n|n-1)\mathbf{u}\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)}{1+\delta_e^2(n|n-1)\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbf{d}} \\ &= \frac{\mathbf{R}_{ee}(n|n-1)\mathbf{u}}{[\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbf{d}]^{-1} + \delta_e^2(n|n-1)} \times \frac{\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)}{\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbf{d}}.\end{aligned} \tag{42}$$

Because $V_o(n) = \mathbf{h}_{\mathrm{mvdr}}^H(n)\mathbf{v}(n)$, we have

$$\begin{aligned}\delta_{V_o}^2 &= \mathbf{h}_{\mathrm{mvdr}}^H(n)\mathbf{R}_{vv}(n)\mathbf{h}_{\mathrm{mvdr}}(n) \\ &= \frac{\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbf{d}}{[\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbf{d}]^2} \\ &= [\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbf{d}]^{-1}.\end{aligned} \tag{43}$$

According to (29) and (43), the MKF gain in (42) is further simplified as

$$\mathbf{G}(n) = \mathbf{R}_{ee}(n|n-1)\mathbf{u}[\delta_{V_o}^2 + \delta_e^2(n|n-1)]^{-1}\mathbf{h}_{\mathrm{mvdr}}^H(n). \tag{44}$$

Substituting the $\delta_e^2(n|n-1)$ in (40) into (44),

$$\begin{aligned}\mathbf{G}(n) &= \mathbf{R}_{ee}(n|n-1)\mathbf{u}[\delta_{V_o}^2 + \mathbf{u}^T\mathbf{R}_{ee}(n|n-1)\mathbf{u}]^{-1}\mathbf{h}_{\mathrm{mvdr}}^H(n) \\ &= \tilde{\mathbf{K}}(n)\mathbf{h}_{\mathrm{mvdr}}^H(n),\end{aligned} \tag{45}$$

where

$$\tilde{\mathbf{K}}(n) = \mathbf{R}_{ee}(n|n-1)\mathbf{u}[\delta_{V_o}^2 + \mathbf{u}^T\mathbf{R}_{ee}(n|n-1)\mathbf{u}]^{-1}. \tag{46}$$

In (45), the MKF gain is factorized into a MVDR beamformer and a new STFT-domain gain $\tilde{\mathbf{K}}(n)$. We can notice that the formulation of $\tilde{\mathbf{K}}(n)$ is similar to the single-channel MDKF gain $\mathbf{K}(n)$ in (35). However, it is defined based on $\mathbf{R}_{ee}(n|n-1)$, which is the covariance matrix of the STFT-domain LP estimation error, while $\mathbf{K}(n)$ is defined based on $\mathbf{R}_{\eta\eta}(n|n-1)$, which is the covariance matrix of the modulation-domain LP estimation error.

C. Comparison between MKF and MVDR-MDKF

In this subsection, based on the factorization of the MKF gain, we analyse the relationship between MKF and the MVDR-MDKF.

First we compare the LP stage in the MKF and in the MVDR-MDKF. In the modulation domain, it can be seen from (8) and (34) that, the LP estimation, which is based on the dynamic model of clean speech and the KF state of the previous frame, is the same in the MKF and in the MVDR-MDKF.

However, for the MKF, the modulation-domain LP estimation is transferred into the STFT domain in (9) before incorporating the multichannel noisy observations in the updating step. In contrast, in the MVDR-MDKF, the modulation-domain LP estimate is directly used in (37) to update the state vector, and the STFT-domain estimate is finally obtained by inserting the phase of MVDR output to the updated state vector as in (38). Therefore, it is useful to analyse the relationship between the two ways of updating the state vector.

First consider the state vector updating of MKF in (12). From (45), since $\mathbf{h}_{\mathrm{mvdr}}^H(n)\mathbf{d} = 1$, which is the constraint of the MVDR beamformer optimization problem [7], (12) can be

expressed as

$$\begin{aligned}\hat{\mathbf{x}}_1(n|n) &= \hat{\mathbf{x}}_1(n|n-1) + \mathbf{G}(n)[\mathbf{y}(n) - \mathbf{Q}\hat{\mathbf{x}}_1(n|n-1)] \\ &= \hat{\mathbf{x}}_1(n|n-1) + \tilde{\mathbf{K}}(n)[\mathcal{Z}_1(n) - \mathbf{u}^T\hat{\mathbf{x}}_1(n|n-1)].\end{aligned} \tag{47}$$

It can be seen that (47) is similar to the single-channel MDKF counterpart (37), but is formulated in the STFT domain. Since $\hat{\mathbf{x}}_1(n|n-1)$ has the same phase as the MVDR output $\mathcal{Z}_1(n)$ as is shown in (9), we have

$$\begin{aligned}\hat{\mathbf{x}}_1(n|n) = \hat{\mathbf{\Phi}}(n)\hat{\mathbf{a}}_1(n|n-1) + \tilde{\mathbf{K}}(n)\hat{\mathbf{\Phi}}_{1,1}(n)[|\mathcal{Z}_1(n)| \\ - \mathbf{u}^T\hat{\mathbf{a}}_1(n|n-1)].\end{aligned} \tag{48}$$

Now we calculate the $\tilde{\mathbf{K}}(n)\hat{\mathbf{\Phi}}_{1,1}(n)$ in (48). From (46), since $\hat{\mathbf{\Phi}}(n)\hat{\mathbf{\Phi}}^H(n) = \mathbf{I}$, it can be deduced that

$$\begin{aligned}&\tilde{\mathbf{K}}(n)\hat{\mathbf{\Phi}}_{1,1}(n) \\ &= \mathbf{R}_{ee}(n|n-1)\mathbf{u}\hat{\mathbf{\Phi}}_{1,1}(n) \times [\delta^2_{V_o} + \mathbf{u}^T\mathbf{R}_{ee}(n|n-1)\mathbf{u}]^{-1} \\ &= \hat{\mathbf{\Phi}}(n)\hat{\mathbf{\Phi}}^H(n)\mathbf{R}_{ee}(n|n-1)\hat{\mathbf{\Phi}}(n)\mathbf{u} \\ &\quad \times [\delta^2_{V_o} + \mathbf{u}^T\mathbf{R}_{ee}(n|n-1)\mathbf{u}]^{-1}.\end{aligned} \tag{49}$$

To further simplify (49), here we use the phase of the MVDR output to approximate the clean phase, and rewrite $\tilde{\mathbf{K}}(n)\hat{\mathbf{\Phi}}_{1,1}(n)$ according to (26),

$$\begin{aligned}&\tilde{\mathbf{K}}(n)\hat{\mathbf{\Phi}}_{1,1}(n) \\ &= \hat{\mathbf{\Phi}}(n)\mathbf{R}_{\eta\eta}(n|n-1)\mathbf{u}[\delta^2_{V_o} + \mathbf{u}^T\mathbf{R}_{ee}(n|n-1)\mathbf{u}]^{-1}.\end{aligned} \tag{50}$$

It can be verified that $\mathbf{u}^T\mathbf{R}_{ee}(n|n-1)\mathbf{u} = \mathbf{u}^T\mathbf{R}_{\eta\eta}(n|n-1)\mathbf{u}$, then we can finally simplify $\tilde{\mathbf{K}}(n)\hat{\mathbf{\Phi}}_{1,1}(n)$ as

$$\begin{aligned}&\tilde{\mathbf{K}}(n)\hat{\mathbf{\Phi}}_{1,1}(n) \\ &= \hat{\mathbf{\Phi}}(n)\mathbf{R}_{\eta\eta}(n|n-1)\mathbf{u}[\delta^2_{V_o} + \mathbf{u}^T\mathbf{R}_{\eta\eta}(n|n-1)\mathbf{u}]^{-1} \\ &= \hat{\mathbf{\Phi}}(n)\mathbf{K}(n).\end{aligned} \tag{51}$$

Substituting (51) into (48), we have

$$\begin{aligned}\hat{\mathbf{x}}_1(n|n) = \hat{\mathbf{\Phi}}(n)\{\hat{\mathbf{a}}_1(n|n-1) + \mathbf{K}(n)[|\mathcal{Z}_1(n)| \\ - \mathbf{u}^T\hat{\mathbf{a}}_1(n|n-1)]\}.\end{aligned} \tag{52}$$

Comparing between (52) and (37), it is evident that, by using the method in Section IV-C to estimate the $\mathbf{R}_{ee}(n|n-1)$, the state update in (12) can be seen as first updating the state vector in the modulation domain using (37), and then inserting the phase of MVDR output $\hat{\mathbf{\Phi}}(n)$ into the updated state vector.

After updating the state vector, now we compare the updating of covariance matrices of the estimation error in both frameworks. Before updating the state vector, it can be viewed that the updating in (21) and (33) are the same. After obtaining the new state vector, because $\mathbf{G}(n)\mathbf{Q} = \tilde{\mathbf{K}}(n)\mathbf{h}^H_{\mathrm{mvdr}}(n)\mathbf{d}\mathbf{u}^T = \tilde{\mathbf{K}}(n)\mathbf{u}^T$, according to (51), the updating of $\mathbf{R}_{\eta\eta}(n|n)$ in (26) is reformulated as

$$\begin{aligned}\mathbf{R}_{\eta\eta}(n|n) &= \hat{\mathbf{\Phi}}^H(n)\mathbf{R}_{ee}(n|n)\hat{\mathbf{\Phi}}(n) \\ &= \hat{\mathbf{\Phi}}^H(n)[\mathbf{I} - \tilde{\mathbf{K}}(n)\mathbf{u}^T]\mathbf{R}_{ee}(n|n-1)\hat{\mathbf{\Phi}}(n) \\ &= \hat{\mathbf{\Phi}}^H(n)[\mathbf{I} - \tilde{\mathbf{K}}(n)\mathbf{u}^T]\hat{\mathbf{\Phi}}(n)\mathbf{R}_{\eta\eta}(n|n-1) \\ &= [\mathbf{I} - \mathbf{K}(n)\mathbf{u}^T]\mathbf{R}_{\eta\eta}(n|n-1),\end{aligned} \tag{53}$$

which is equivalent to the updating in (36).

In the following, we discuss the optimality of MKF and MVDR-MDKF.

For further discussion, here we first distinguish between two versions of MKF in Section IV: a) the *optimal MKF* with the MKF gain as in (18); b) the *practical MKF* whose $\mathbf{R}_{ee}(n|n-1)$ is estimated using the method in Section IV-C. As the optimal MKF is derived under the MMSE criterion using the STFT-domain error signal $\mathbf{e}(n|n)$, theoretically, it always provides both correct amplitude and phase estimates. However, since $\mathbf{R}_{ee}(n|n-1)$ is unknown in practice, the MKF is implemented based on the proposed $\mathbf{R}_{ee}(n|n-1)$ estimation method in Section IV-C. We notice that the $\mathbf{R}_{ee}(n|n-1)$ estimation uses the phase of the MVDR output to approximate the clean phase, therefore, if the approximation is not accurate in practice, the practical MKF will not yield STFT-domain optimal estimation of the clean speech. We note that if better $\mathbf{R}_{ee}(n|n-1)$ estimation method exists, with the formulations of optimal MKF, the STFT-domain clean speech can be more accurately estimated.

On the other hand, by using the phase of the MVDR output to approximate the clean phase, the relationships in (22) and (26) can be derived. In Section IV-C and this section, the relationships in (22) and (26) are used both for $\mathbf{R}_{ee}(n|n-1)$ estimation, and for the derivations from (50) to (53). As a result, we can conclude that, as long as the practical MKF and MVDR-MDKF adopt the same state transition matrix $\mathbf{B}$, and the noise variance used in MDKF is calculated by (43), the practical MKF and MVDR-MDKF will always yield the same results.

From the above analysis, we can further infer that, if the clean phase of speech cannot be accurately approximated by the phase of the MVDR output, both the practical MKF and MVDR-MDKF are unable to give the STFT-domain optimal estimation of the clean signal. However, by factorizing the practical MKF into MVDR-MDKF in (52), we have decoupled the amplitude estimation and phase estimation of the clean signal. If the phase of the MVDR output is correct, both algorithms will yield the same and optimal estimation. It can be seen from Section V-A that the amplitude estimation of MVDR-MDKF does not rely on the phase information. Thus we can deduce that, even when the phase of the MVDR output fails to accurately approximate the clean phase, the estimated amplitude of MVDR-MDKF remains unchanged, therefore, the MVDR-MDKF and practical MKF, always provide optimal amplitude estimations, in the multichannel case under the MMSE criterion.

D. Using MVDR-MDKF as an Alternative Implementation of MKF

Based on the above analysis, it is natural to propose the MVDR-MDKF as an alternative implementation of the MKF.

TABLE II
NORMALIZED EXECUTION TIME OF DIFFERENT ALGORITHMS

Algorithm	MVDR	MWF	MKF	MVDR-MDKF
Execution time	1	1.58	4.79	1.82

TABLE III
AZIMUTHS OF DIFFERENT SOURCE POSITIONS

Position	1_A	1_B	1_C	1_D	1_E
Azimuth	0° ↑	27° ↖	90° ←	-90° →	-140° ↘

Compared with the MKF in Section IV, the main advantage of this alternative implementation lies in the computational efficiency. Although the MKF and MVDR-MDKF adopt the same strategy in the modulation LP, for the MKF, the modulation-domain LP estimate is further converted into the STFT domain using (9). To compute the optimal MKF gain, (17) involves multiplications between the $M \times P$ matrix $\mathbf{Q}$ and $\mathbf{R}_{ee}(n|n-1)$, and the matrix inversion. In contrast, since $\mathbf{u}$ is a vector, computing the MDKF gain from (35) mainly involves extracting one column of $\mathbf{R}_{\eta\eta}(n|n-1)$ and the division by a scalar. Moreover, both $\mathbf{R}_{\eta\eta}(n)$ and $\mathbf{R}_{ee}(n)$ need to be updated in the MKF, while MVDR-MDKF only updates $\mathbf{R}_{\eta\eta}(n)$ in the modulation domain.

To evaluate the computational complexity of different algorithms, we run each algorithm for 50 trials, and the average execution time of each algorithm is shown in Table II. The execution time is normalized with respect to that of the MVDR. The computational time for noise covariance matrix estimation is not counted. Since the MKF and MVDR-MDKF additionally uses MWF to obtain the pre-processed signal for LP analysis, the computational complexity of MKF and MVDR-MDKF is higher than MWF as well as MVDR. However, compared with MKF, the computational complexity of MVDR-MDKF is largely reduced.

Although the idea of concatenating the MVDR with the single-channel MDKF seems simple, based on the analysis and discussions in Section V-C, we show the MVDR-MDKF is equivalent to the MKF in Section IV, which reveals the necessity of the concatenation, and its optimality in the multichannel MMSE sense. On the other hand, the advantage of MKF over MWF is more intuitively shown as the MWF can be factorized as MVDR and single-channel Wiener filter [13], and it has been demonstrated that the single-channel MDKF can generally perform better than single-channel Wiener filter.

VI. EXPERIMENTS

To evaluate the performances of different speech enhancement algorithms, in this section we conduct experiments using an HRIR hearing aid (HA) database [48] measured in real reverberant environments.

A. Experimental Setup

We utilize the eight-channel HRIR database from [48] to generate multichannel signals at a sampling frequency of 8 kHz. The eight channels include one in-ear channel and three behind-the-ear (BTE) channels for each ear. Real RIRs and noises are recorded in different types of environment including a cafeteria, office and courtyard. We will mainly test the cafeteria environment which is reverberant and typically more noisy than the office and courtyard environments, although comparison results for different environments will also be presented. In all experiments, spatially white Gaussian noise is not further added to account for the sensor noise, since the effect of sensor noise is already included in the measured RIRs and ambient noises.

A 10 s speech source signal is generated by concatenating randomly selected sentences from the IEEE sentences database [63], and this is then convolved with measured RIRs to generate the clean reverberant speech signals in multiple microphones. Algorithms are tested both with and without an interfering source, and details of noise signals and sound source locations will be described in each experiment. Table III summarizes the source position azimuths that are considered in the experiments, where 0° is straight in front of the listener and 90° is to their left.

We test four algorithms which include the proposed MKF in Section IV and its efficient implementation MVDR-MDKF in Section V-A, as well as the MVDR and MWF. As has been discussed in the introduction, the MWF can be implemented by either only relying the noise covariance matrix as in [21], or as a concatenation of MVDR and single-channel Wiener post-filter based on the knowledge of RTF as in [55]. Since the proposed methods and MVDR exploit the RTF, for fair comparison, the post-filtering structure of the MWF is adopted. For all algorithms, the STFT window is 16 ms with 4 ms frame hop. The RTF vector is computed using 16 ms truncated RIRs, and the first channel is taken as the reference. We estimate the multichannel noise covariance matrix using the method in [56]. In the MKF and MVDR-MDKF, we set the LPC order $P = 2$ and, to estimate the LP coefficients $a_{p,k}$ and the excitation variance $W(n)$ in (4), the modulation frame duration is 32 ms with 16 ms frame hop. Therefore, eight STFT frames are included in one modulation frame to perform LP analysis, and the LP coefficients are updated only every four STFT frames. The parameters of the pre-processing MWF used in MKF and MVDR-MDKF are the same as those of the baseline MWF.

B. Performance Measure

Three metrics are used to evaluate the speech enhancement performance, which include the short-time objective intelligibility (STOI) [64], perceptual evaluation of speech quality (PESQ) [65], and frequency-weighted segmental signal-to-noise ratio (FwSegSNR) [66]. For each metric, both the raw value of the reference noisy signal ("$[\cdot]_{\mathrm{raw}}$") and the improvement ("$\Delta[\cdot]$") in the output signal are computed. The first two seconds of the signals are excluded from the evaluation to allow for the convergence of noise estimation algorithms. The metrics of the noisy input and the enhanced signal are averaged over ten trials with different randomly chosen speech signals.

For each set of experiments, a t-test is also performed to determine whether the proposed MKF and MVDR-MDKF are

TABLE IV
THE TABLE GIVES P VALUES FOR PAIRED T-TESTS COMPARING THE PERFORMANCE OF THE PROPOSED MKF ALGORITHMS WITH THE MVDR AND MWF ALGORITHMS

		ΔSTOI		ΔPESQ		ΔFwSegSNR	
		MVDR	MWF	MVDR	MWF	MVDR	MWF
Effect of SNR	Ambient Noise	0.770	0.469	**<0.001**	**<0.001**	**<0.001**	**<0.001**
	Babble Noise	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
Effect of Number of Microphones	Ambient Noise	0.417	0.503	**<0.001**	**<0.001**	**<0.001**	**<0.001**
	Babble Noise	0.925	0.852	**<0.001**	**<0.001**	**<0.001**	**<0.001**
Effect of SIR	SSN	0.347	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
	WGN	0.289	0.423	**<0.001**	**<0.001**	**<0.001**	**<0.001**
Effect of Interference Position	SSN	**0.009**	0.433	**<0.001**	**<0.001**	**<0.001**	**<0.001**
	WGN	0.060	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
Effect of Environment Type	SSN	0.135	0.380	**<0.001**	**<0.001**	**<0.001**	**<0.001**
	WGN	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**

Comparisons that are Significant at the 2.5% Individual Level (with 5% overall level and Bonferroni correction) are Shown in Bold.

significantly different from the baseline methods. As has been discussed in Section V-C and will be shown later in the experiments, the MKF and MVDR-MDKF always yield identical results, therefore for each baseline algorithm, only one t-test result is presented. The overall significance level is set to be 5% in all tests. As two hypotheses are tested, Bonferroni correction is used, which lead to a 2.5% individual significance level. To facilitate the discussion in the following subsections, we summarize in advance the results in Table IV.

C. Experimental Results Without Interfering-Sources

In this subsection, we test the scenario with a single target speaker and no interfering sources. The experiments are conducted in a cafeteria environment in which the speaker is seated at 0° azimuth in front of the listener (position A in Fig. 5 of [67]), and the speaker and listener are seated at a rectangle table placed near one corner of the room. Diffuse ambient noise and babble noise is added to the clean signal, which were recorded during the off-peak and peak times, respectively.

1) Effect of SNR: We first study algorithm performance as a function of SNR. Since the SNR varies across channels, we determine the scaling factor of the additive noise using the first channel as reference, and the first channel is chosen arbitrarily as the left BTE front microphone. The multichannel background noise is then scaled and added to the clean reverberant speech signal. The tested SNRs range from −5 dB to 15 dB in 5 dB increments, and six channels, which exclude the two in-ear channels, are used in the experiments.

The comparison results are shown in Fig. 2, in which the columns are for ambient and babble noise respectively and the rows show different metrics. On each graph, the dashed line plots the raw metric value for the noisy reference channel using the right-side axis while the bars show deviations from these values using the left-side axis. Firstly, it can be seen that the MKF and its efficient implementation MVDR-MDKF always yield identical results, which is consistent with the theoretical analysis in Section V-C. The Table IV gives P values for paired t-tests comparing the performance of the proposed MKF algorithms with the MVDR and MWF algorithms. Comparisons that are significant at the 5% level are shown in bold. Fig. 2(a)(b) show that all the algorithms yield very similar STOI

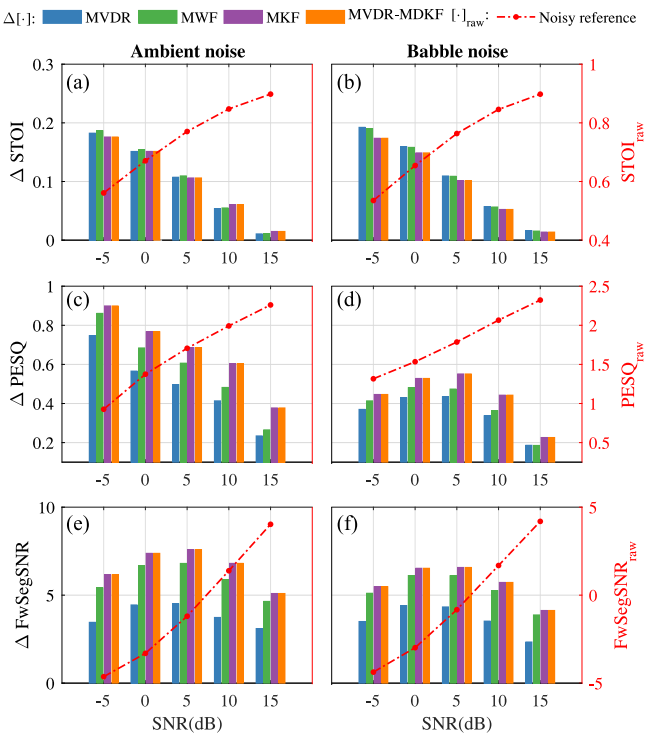

Fig. 2. The comparison results in different SNRs without interfering source in the cafeteria environment. Six channels are used. The performance improvements ("Δ[·]") and the raw values ("$[\cdot]_{\mathrm{raw}}$") of the reference noisy signal are shown using bars and dashed lines, respectively.

performance although the MKF algorithms are slightly worse at low SNRs. From Table IV, this difference is significant (at the 5% level) for babble noise but not for ambient noise. The STOI measure estimates the intelligibility from the modulation features of speech, so it might be expected that the MKF would perform well using this measure. A possible reason for the meagre STOI improvement is that the frame lengths used for computing STOI and MKF are different. Whereas the frame length for computing STOI is 384 ms [64], the LP modelling in our MKF has a support of only 32 ms.

In contrast, Table IV and Fig. 2(c) to (e) show that, the proposed methods yield significantly greater improvements in PESQ and FwSegSNR than the MVDR and MWF algorithms.

TABLE V
INFORMAL WEBMUSHRA TEST RESULTS

Description	Reference	Noisy	MVDR	MWF	Proposed
Average score	100	21.35	52.81	63.63	72.28

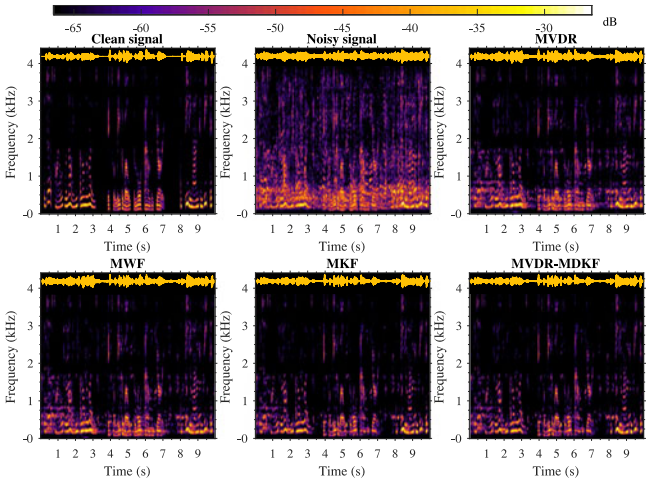

Fig. 3. Clean, noisy and enhanced spectrograms of different algorithms in the 0 dB babble noise.

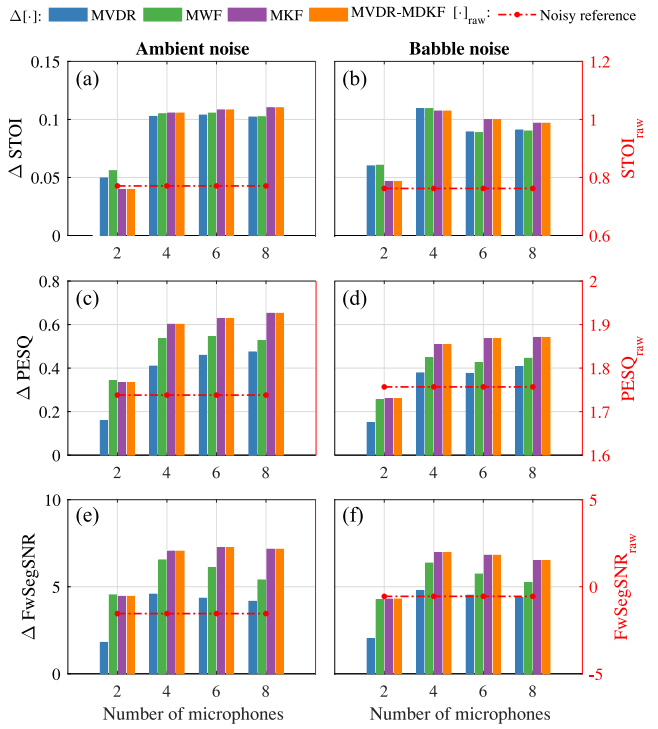

Fig. 4. The comparison results using different number of microphones without interfering source in the cafeteria environment. The SNR is fixed to 5 dB. The performance improvements ("$\Delta[\cdot]$") and the raw values ("$[\cdot]_{\mathrm{raw}}$") of the reference noisy signal are shown using bars and dashed lines, respectively.

On average, compared with MVDR and MWF respectively, the improvements of the proposed methods in the ambient noises are 0.2 and 0.17 higher in PESQ, and are 2.8 and 0.7 dB higher in FwSegSNR. For the babble noise cases, these values become 0.1, 0.06, 2.1 and 0.4, respectively.

It can be seen that the improvements in babble noise noise (right column of Fig. 2) are smaller for the ambient noise (right column). A possible reason for this is that estimating the less stationary babble noise is more difficult. In addition, generally the improvements of different metrics decrease when the SNR increases. This is because in such conditions the difference between the clean and noisy speech becomes smaller, so that there is less potential for improvement.

Informal subjective listening test for the scenarios considered in this subsection is also conducted using the webMUSHRA test [68]. The 12 listeners were asked to rate the overall quality of different enhanced signals, the reference signal as well as the noisy signal. Eight cases are tested, which include one male or one female speaker, the ambient noise or babble noise, and the -5 dB or 5 dB SNR, respectively. The results are shown in Table V and also demonstrate the effectiveness of the proposed MKF and MVDR-MDKF.

Fig. 3 shows an example of the clean, noisy and enhanced spectrograms of different algorithms in the babble noise at 0 dB SNR. It can be seen that, the MWF yields less noisy output than the MVDR and, by exploiting the temporal evaluation of the speech signal, the residual noise is further reduced by MKF and MVDR-MDKF.

2) Effect of Number of Microphones: In this set of experiments, we fix the SNR of the reference microphone to be 5 dB, and use between two and eight microphones for speech enhancement. According to [48], the eight channels in the HRIR database include the three pairs of BTE microphones which are located in the front, middle and back of the ears ears respectively, and one pair of in-ear microphones. We increase the number of microphones by first only using the BTE front microphone pair, and then successively including the BTE middle, back and in-ear microphone pairs.

The results are given in Fig. 4. When using only two microphones, it can be observed that the proposed methods cannot achieve better performances than MWF. Since the MKF-based methods rely on the information of dynamic model, which is estimated using the MWF output, if only two microphones are used, the noise residual in the MWF output makes the estimated dynamic model inaccurate, and further degrades the performances of the MKF-based methods. When using four microphones, the performance of all methods increases significantly, however, with diminishing improvements towards 8 microphones. Similarly to the previous subsection, it is shown that the proposed methods yield the largest improvements in PESQ and FwSegSNR, and have comparable STOI improvements with conventional methods when at least four microphones are used.

D. Experimental Results With Interfering-Sources

In practical scenarios, directional interference sources may exist in addition to diffuse environmental noise. To test the performance under interfering source conditions, first we consider the cafeteria environment but with an additional interfering source. The influence of signal-to-interference ratio (SIR) and

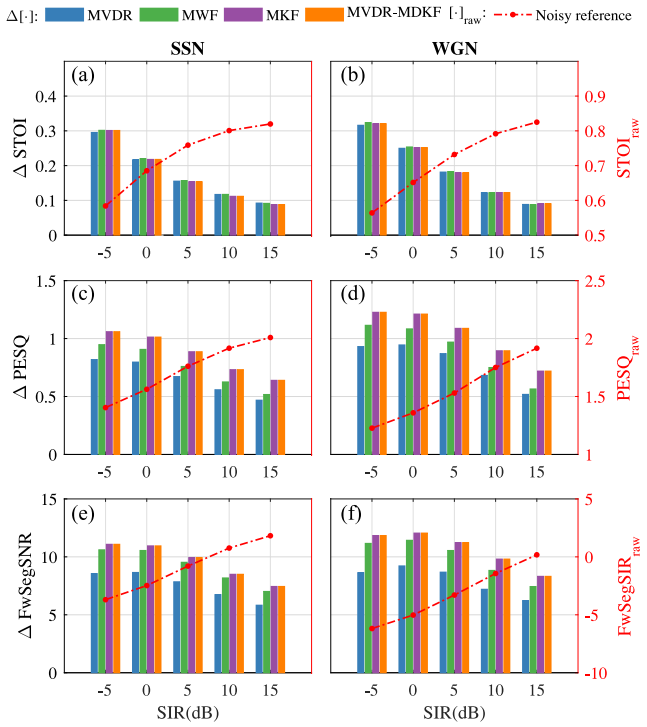

Fig. 5. The comparison results in different SIRs with interfering source present in the cafeteria environment. 10 dB SNR babble noise is present in all cases. Six channels are used. The performance improvements ("$\Delta[\cdot]$") and the raw values ("$[\cdot]_{\mathrm{raw}}$") of the reference noisy signal are shown using bars and dashed lines, respectively.

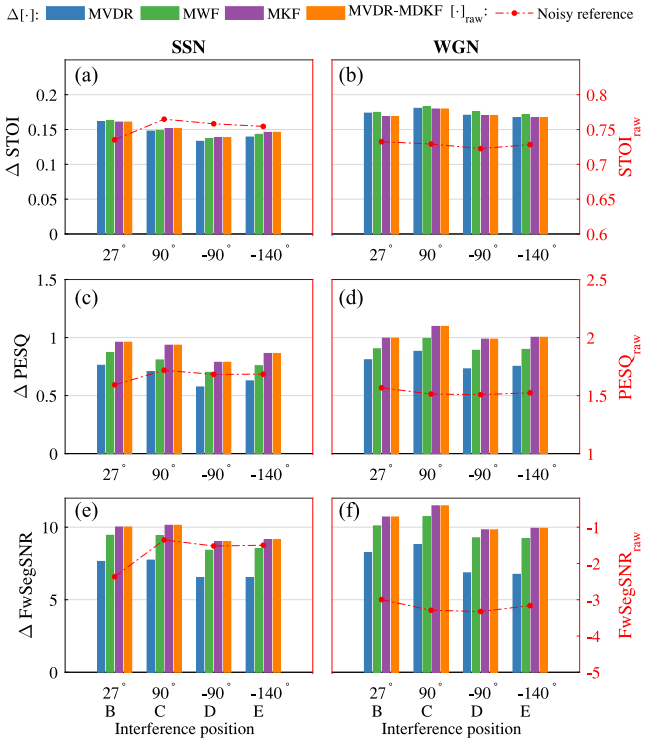

Fig. 6. The comparison results for different interfering source positions in the cafeteria environment. The SIR is 5 dB, and 10 dB SNR babble noise is present in all cases. Six channels are used. The performance improvements ("$\Delta[\cdot]$") and the raw values ("$[\cdot]_{\mathrm{raw}}$") of the reference noisy signal are shown using bars and dashed lines, respectively.

the interfering source position on the performance of different algorithms is evaluated. Second, we conduct experiments in different environment types (cafeteria, office and courtyard), and present the average performance of the different methods.

1) Effect of SIR: In this set of experiments, the target speaker is located at 90° azimuth and a one single interfering source is always located to the left of the listener (position C in Fig. 5 of [67]) emitting directional speech shaped noise (SSN) or white Gaussian noise (WGN). The SIR changes from −5 dB to 15 dB in 5 dB increments, and is evaluated using the left BTE front microphone as reference. To simulate a more realistic environment, 10 dB SNR environmental babble noise is additionally added to the noisy and reverberant speech. Again, six channels which exclude the two in-ear channels are used.

The performance is shown in Fig. 5, in which the columns correspond to an interfering noise source that is SSN or WGN respectively. When SIR ≤ 0 dB, while achieving similar improvements to conventional methods in STOI, from Fig. 5(c) to (e), the proposed methods can obtain at least 1 unit improvement in PESQ and 10 dB improvement in FwSegSNR, which are the highest among all methods. When SIR > 0 dB, the improvements of all algorithms tend to be lower, nevertheless, the proposed methods always attain the best performances in Δ PESQ and Δ FwSegSNR.

It can be also noticed that, for PESQ, in both SSN and WGN cases, the performance differences between the proposed methods and MWF are close to the PESQ differences of the reference noisy signal when increasing the SIR by 5 dB, indicating that the proposed methods have a gain of approximate 5 dB over MWF in PESQ.

The methods generally achieve larger improvements for WGN than for SSN, because SSN has spectral properties more similar to those of the target speech.

2) Effect of Interfering Source Position: We further investigate the influence of the interfering source position on speech enhancement performance. In the cafeteria environment, with the target speaker seated in front of the listener at position 1_A, we test four interfering source positions which include the 1_{B to E}, corresponding to the azimuths from 27°, 90°, −90°, −140°, respectively (Table III). We consider the cases for which the interfering signal is the SSN or WGN. In all cases the SIR is fixed at 5 dB and 10 dB SNR environmental babble noise is always present. We use six channels excluding the two in-ear channels as the inputs to the algorithms.

In Fig. 6 we summarize the results for different interfering source positions. Again, it can be seen that, for different interfering source positions, the MKF and MVDR-MDKF always have larger improvements in PESQ and FwSegSNR. For STOI, according to Table IV, the proposed methods generally perform slightly better than MVDR in SSN, and slightly worse than MWF in WGN.

In general, for both SSN and WGN interference, the algorithms perform worse when the interfering source is located at 1_D (−90°) and 1_E (−140°) than at 1_B (27°) and 1_C

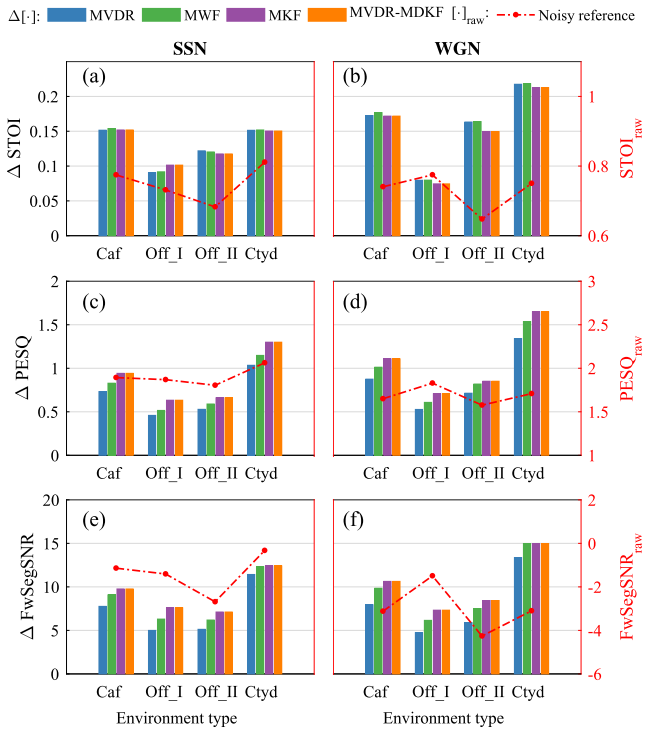

Fig. 7. The comparison results for environment types. The SIR is 5 dB, six channels are used. Caf, Off_I, Off_II and Ctyd stand for cafeteria, office_I, office_II and courtyard. The performance improvements ("$\Delta[\cdot]$") and the raw values ("$[\cdot]_{\mathrm{raw}}$") of the reference noisy signal are shown using bars and dashed lines, respectively.

(90°). From Section VI-D 1), the SIR is evaluated by taking the left BTE front microphone as reference, and the position 1_D (−90°) and 1_E (−140°) are located at the right of the listener. Due to the shadowing effect of the head, for a certain SIR at the left-side reference microphone, the noise level at right-side microphones is higher when the interfering source is on the right than that when the interfering source is on the left. As a consequence, the performance of different algorithms tends to degrade when the interfering source is at 1_D (−90°) and 1_E (−140°).

3) Effect of Environment Type: Finally, we examine the behaviour of speech enhancement algorithms in different environment types. Four environment types including the cafeteria, office I, office II and courtyard are tested. For each environment type, we randomly choose two different RIRs to represent the responses from the target source and interfering source respectively, and the two RIRs correspond to the same head orientation of the listener. We fix the SIR to 5 dB while no ambient environmental noise is added, since the environmental noise changes over environment types. Again, SSN and WGN interference types are tested. We used 50 random combinations of the source and interference positions, and present the average of the metrics of different combinations to obtain the final evaluation of the environment type.

In Fig. 7(a)(b) and Table IV, for different environment types, again, there are no significant differences between the algorithms in terms of STOI performance. In Fig. 7(c)(d), the proposed methods outperform the MVDR and MWF in Δ PESQ, and the improvements are larger when the interfering source is WGN. For the FwSegSNR, it can be seen that the proposed methods perform best in cafeteria and office conditions, and have similar performances to MWF in courtyard conditions. In addition, it is shown that all the methods achieve the largest improvements in the courtyard conditions, and the least improvements in the office conditions.

VII. Conclusion

A modulation-domain MKF is proposed in this paper for multichannel speech enhancement. The key feature of the MKF is that it exploits spatial information as well as the temporal information simultaneously to estimate the clean speech signal. To use the spatial and temporal information jointly, the MKF performs LP in the modulation domain and incorporates the spatial information in the STFT domain. An optimal MKF gain is derived to adaptively combine the LP estimation and observation. We show that the MKF is equivalent to the MWF when the LP estimate is not used. We further show that the MKF can be factorized as a MVDR followed by a MDKF post-filter, thus an alternative implementation of the MKF, MVDR-MDKF, is proposed. The experiments show that the MKF and the alternative implementation MVDR-MDKF always give the same results and outperform conventional MVDR and MWF in various noisy and reverberant conditions.

References

[1] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 549–557, Nov. 2003.
[2] J. Benesty, J. Chen, Y. Huang, and J. Dmochowski, "On microphone-array beamforming from a MIMO acoustic signal processing perspective," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1053–1065, Mar. 2007.
[3] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 260–276, Feb. 2010.
[4] E. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the MVDR beamformer in room acoustics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 158–170, Jan. 2010.
[5] W. Herbordt and W. Kellermann, "Adaptive beamforming for audio signal acquisition," in *Adaptive Signal Processing: Applications to Real-World Problems*, (Signals and Communication Technology), J. Benesty and Y. Huang, Eds., Berlin, Germany: Springer-Verlag, 2003, ch. 6, pp. 155–194.
[6] T. Yu and J. Hansen, "A speech presence microphone array beamformer using model based speech presence probability estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 213–216.
[7] C. Pan, J. Chen, and J. Benesty, "Performance study of the MVDR beamformer as a function of the source incidence angle," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 1, pp. 67–79, Jan. 2014.
[8] M. Souden, J. Benesty, and S. Affes, "A study of the LCMV and MVDR noise reduction filters," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4925–4935, Sep. 2010.
[9] J. Chen, J. Benesty, and C. Pan, "On the design and implementation of linear differential microphone arrays," *J. Acoust. Soc. Amer.*, vol. 136, no. 6, pp. 3097–3113, Dec. 2014.
[10] B. Widrow, P. Mantey, L. Griffiths, and B. Goode, "Adaptive antenna systems," *Proc. IEEE*, vol. 55, no. 12, pp. 2143–2159, Dec. 1967.
[11] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.
[12] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, Jul. 1996.

[13] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001.

[14] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, New York, NY, USA, Apr. 1988, pp. 2578–2581.

[15] B. Cauchi *et al.*, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP J. Adv. Signal Process.*, vol. 2015, no. 1, pp. 1–12, 2015.

[16] I. McCowan and H. Bourlard, "Microphone array post-filter for diffuse noise field," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2002, vol. 1, pp. 905–908.

[17] I. Cohen and B. Berdugo, "Microphone array post-filtering for non-stationary noise suppression," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Orlando, FL, USA, May 2002, pp. 901–904.

[18] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[19] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. AP-30, no. 1, pp. 27–34, Jan. 1982.

[20] B. R. Breed and J. Strauss, "A short proof of the equivalence of LCMV and GSC beamforming," *IEEE Signal Process. Lett.*, vol. 9, no. 6, pp. 168–169, Jun. 2002.

[21] 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.

[22] S. Doclo, A. Spriet, and M. Moonen, "Efficient frequency-domain implementation of speech distortion weighted multi-channel Wiener filtering for noise reduction," in *Proc. Eur. Signal Process. Conf.*, 2004, pp. 2007–2010.

[23] S. Doclo and M. Moonen, "On the output SNR of the speech-distortion weighted multichannel Wiener filter," *IEEE Signal Process. Lett.*, vol. 11, no. 12, pp. 809–811, Dec. 2005.

[24] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Process.*, vol. 84, no. 12, pp. 2367–2387, Dec. 2004.

[25] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds., Berlin, Germany: Springer-Verlag, 2001, ch. 3, pp. 39–60.

[26] K. K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Dallas, TX, USA, Apr. 1987, pp. 177–180.

[27] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 373–385, Jul. 1998.

[28] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Signal Process.*, vol. 39, no. 8, pp. 1732–1742, Aug. 1991.

[29] E. Zavarehei, S. Vaseghi, and Q. Yan, "Speech enhancement with Kalman filtering the short-time DFT trajectories of noise and speech," in *Proc. IEEE Eur. Signal Process. Conf.*, 2006, pp. 1–5.

[30] T. Esch and P. Vary, "Speech enhancement using a modified Kalman filter based on complex linear prediction and supergaussian priors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2008, pp. 4877–4880.

[31] S. So and K. K. Paliwal, "Modulation-domain Kalman filtering for single-channel speech enhancement," *Speech Commun.*, vol. 53, no. 6, pp. 818–829, Jul. 2011.

[32] Y. Wang and M. Brookes, "Speech enhancement using a robust Kalman filter post-processor in the modulation domain," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 7457–7461.

[33] Y. Wang and M. Brookes, "Speech enhancement using a modulation domain Kalman filter post-processor with a Gaussian mixture noise model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 7024–7028.

[34] Y. Wang and M. Brookes, "Speech enhancement using an MMSE spectral amplitude estimator based on a modulation domain Kalman filter with a Gamma prior," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5225–5229.

[35] N. Dionelis and M. Brookes, "Modulation-domain speech enhancement using a Kalman filter with a Bayesian update of speech and noise in the log-spectral domain," in *Proc. Joint Workshop Hands-Free Speech Commun. Microphone Arrays*, 2017, pp. 111–115.

[36] N. Dionelis and M. Brooke, "Speech enhancement using modulation-domain Kalman filtering with active speech level normalized log-spectrum global priors," in *Proc. Eur. Signal Process. Conf.*, 2017, pp. 2309–2313.

[37] Y. Wang, "Speech enhancement in the modulation domain," Ph.D. dissertation, Department of Electrical and Electronic Engineering, Imperial College London, London, U.K., 2015.

[38] Y. Wang and M. Brookes, "Model-based speech enhancement in the modulation domain," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 3, pp. 580–594, Mar. 2018.

[39] N. Dionelis and M. Brookes, "Phase-aware single-channel speech enhancement with modulation-domain Kalman filtering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 5, pp. 937–950, May 2018.

[40] N. Mesgarani and S. Shamma, "Speech enhancement based on filtering the spectrotemporal modulations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2005, vol. 1, pp. 1105–1108.

[41] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, 1994, Art no. 1053.

[42] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

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

[44] S. Braun and E. A. Habets, "Online dereverberation for dynamic scenarios using a Kalman filter with an autoregressive model," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1741–1745, Dec. 2016.

[45] J. Traa and P. Smaragdis, "A wrapped Kalman filter for azimuthal speaker tracking," *IEEE Signal Process. Lett.*, vol. 20, no. 12, pp. 1257–1260, Dec. 2013.

[46] D. Bechler, M. Grimm, and K. Kroschel, "Speaker tracking with a microphone array using Kalman filtering," *Adv. Radio Sci.*, vol. 1, pp. 113–117, May 2003.

[47] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME J. Basic Eng.*, vol. 82, no. Series D, pp. 35–45, 1960.

[48] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP J. Adv. Signal Process.*, vol. 2009, 2009, Art no. 298605.

[49] W. Xue, A. H. Moore, M. Brookes, and P. A. Naylor, "Multichannel Kalman filtering for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.,* 2018, pp. 41–45.

[50] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, Sep. 2004.

[51] T. Esch and P. Vary, "Exploiting temporal correlation of speech and noise magnitudes using a modified Kalman filter for speech enhancement," in *Proc. Voice Commun. (SprachKommunikation) ITG Conf.*, 2008, pp. 1–4.

[52] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. Berlin, Germany: Springer-Verlag, 2010.

[53] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[54] Y. Ephraim and D. Mala, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[55] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1595–1608, Sep. 2016.

[56] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2159–2169, Sep. 2011.

[57] R. Hendriks and T. Gerkmann, "Noise correlation matrix estimation for multi-microphone speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 223–233, Jan. 2012.

[58] M. Taseska and E. A. P. Habets, "MMSE-based blind source extraction in diffuse noise fields using a complex coherence-based a priori SAP estimator," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Sep. 2012, pp. 1–4.

[59] M. Brookes, "Differentials of trace," Imperial College London, Website, 1998–2017. [Online]. Available: http://www.ee.imperial.ac.uk/hp/staff/dmb/matrix/calculus.html#deriv_tr ace

[60] R. Balan and J. Rosca, "Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase," in *Proc. Sensor Array Multichannel Signal Process. Workshop*, Aug. 2002, pp. 209–213.

[61] J. Capon, "High resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.
[62] M. Brookes, "Inversion identities," Imperial College London, Website, 1998–2017. [Online]. Available: http://www.ee.imperial.ac.uk/hp/staff/dmb/matrix/identity.html#InvLemma
[63] E. H. Rothauser *et al.*, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. AU-17, no. 3, pp. 225–246, Sep. 1969.
[64] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[65] *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, International Telecommunications Union (ITU-T) Recommendation, Geneva, Switzerland, P.862, Feb. 2001.
[66] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.
[67] R. M. Baumgärtel *et al.*, "Comparing binaural pre-processing strategies I: Instrumental evaluation," *Trends Hearing*, vol. 19, pp. 1–16, 2015.
[68] M. Schoeffler *et al.*, "webMUSHRA—A comprehensive framework for web-based listening tests," *J. Open Res. Softw.*, vol. 6, no. 1, p. 8, 2018.

Wei Xue (M'16) received the B.Eng. degree in automatic control from the Huazhong University of Science and Technology, Wuhan, China, in 2010, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2015. Since August 2015, he is a first Marie Curie experienced Researcher and then a Research Associate with the Department of Electrical and Electronic Engineering, Imperial College London, London, U.K. He was a visiting scholar with Université de Toulon in July 2015, and KU Leuven in September 2016. His research interests are in microphone arrays based speech signal processing, including speech enhancement, sound source localization, and blind system identification.

Alastair H. Moore (M'13) received the M.Eng. degree in electronic engineering in Music Technology Systems from the University of York, York, U.K., in 2005 and the Ph.D. degree from the University of York, York, U.K. He spent three years as a Hardware Design Engineer for Imagination Technologies plc designing digital radio and networked audio consumer electronics products. In 2012, he joined the Department of Electrical and Electronic Engineering, Imperial College London as a Postdoctoral Research Associate. His research interests are in the field of speech and audio processing, especially microphone array signal processing, modeling and characterization of room acoustics, dereverberation and spatial audio perception with applications for robot audition and hearing aids.

Mike Brookes (M'88) is a Reader (Associate Professor) in signal processing with the Department of Electrical and Electronic Engineering, Imperial College London, London, U.K.. After graduating in mathematics from Cambridge University in 1972, he worked with the Massachusetts Institute of Technology and, briefly, the University of Hawaii before returning to the U.K. and joining Imperial College in 1977. Within the area of speech processing, he has concentrated on the modeling and analysis of speech signals, the extraction of features for speech and speaker recognition and on the enhancement of poor quality speech signals. He is the primary author of the VOICEBOX speech processing toolbox for MATLAB. Between 2007 and 2012, he was the Director of the Home Office sponsored Centre for Law Enforcement Audio Research, which investigated techniques for processing heavily corrupted speech signals. He is currently a Principal investigator of the E-LOBES project that seeks to develop environment-aware enhancement algorithms for binaural hearing aids.

Patrick A. Naylor (SM'07) received the B.Eng. degree in electronic and electrical engineering from the University of Sheffield, Sheffield, U.K., and the Ph.D. degree from Imperial College London, London, U.K. He is a member of academic staff with the Department of Electrical and Electronic Engineering, Imperial College London. His research interests are in the areas of speech, audio and acoustic signal processing. He has worked in particular on adaptive signal processing for speech dereverberation, blind multichannel system identification and equalization, acoustic echo control, speech quality estimation and classification, single and multichannel speech enhancement, and speech production modeling with particular focus on the analysis of the voice source signal. In addition to his academic research, he enjoys several fruitful links with industry in the U.K., USA, and in Europe. He is the past-Chair of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing and the Director of the *European Association for Signal Processing*. He was an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS and is currently a Senior Area Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.