Speech Enhancement Based on Modulation-Domain Parametric Multichannel Kalman Filtering

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

***Abstract*—Recently we presented a modulation-domain multichannel Kalman filtering (MKF) algorithm for speech enhancement, which jointly exploits the inter-frame modulation-domain temporal evolution of speech and the inter-channel spatial correlation to estimate the clean speech signal. The goal of speech enhancement is to suppress noise while keeping the speech undistorted, and a key problem is to achieve the best trade-off between speech distortion and noise reduction. In this paper, we extend the MKF by presenting a modulation-domain parametric MKF (PMKF) which includes a parameter that enables flexible control of the speech enhancement behaviour in each time-frequency (TF) bin. Based on the decomposition of the MKF cost function, a new cost function for PMKF is proposed, which uses the controlling parameter to weight the noise reduction and speech distortion terms. An optimal PMKF gain is derived using a minimum mean squared error (MMSE) criterion. We analyse the performance of the proposed MKF, and show its relationship to the speech distortion weighted multichannel Wiener filter (SDW-MWF). To evaluate the impact of the controlling parameter on speech enhancement performance, we further propose PMKF speech enhancement systems in which the controlling parameter is adaptively chosen in each TF bin. Experiments on a publicly available head-related impulse response (HRIR) database in different noisy and reverberant conditions demonstrate the effectiveness of the proposed method.**

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

I. Introduction

SPEECH quality and intelligibility can be severely degraded by additive acoustic noise, and the importance of speech enhancement has been increasingly recognized due to the wide use of speech processing systems such as hearing aids, robotics and smart home devices in noise. Compared with single-channel speech enhancement which relies on the noisy observation from a single microphone, microphone array based multichannel speech enhancement has attracted much attention since the spatial information of the acoustic environment can additionally be exploited to improve performance.

Conventional multichannel speech enhancement methods can be categorized according to whether they are based on beamforming [1]–[6], post-filtering [7]–[10], generalized sidelobe cancelling (GSC) [11], [12], or multichannel Wiener filtering (MWF) [13]–[16]. These methods generally design an optimal filter by exploiting the spatial correlation between multiple microphone signals, in order to estimate a single target signal. Based on the steering vector or the relative transfer function (RTF) which characterises the spatial correlation, fixed beamformers such as delay-and-sum (DS) [17], and adaptive beamformers such as minimum variance distortion response (MVDR) [2], [4], [6] and linearly constrained minimum variance (LCMV) [5], can be viewed as spatial filters that extract the signal from the target direction and attenuate the noise from other directions. In post-filtering [7]–[10] and GSC [11], [12], single-channel noise reduction and multi-channel adaptive noise cancellation (ANC) are used to further reduce the residual noise in the beamforming output. It is shown in [11] that the GSC can be expressed as an unconstrained form of the LCMV beamformer. Although the MWF [13]–[16] does not explicitly rely on knowledge of the steering vector or RTF, it jointly uses the spatial covariance matrices of the noise and mixture signals to implicitly resolve the spatial information of the target. It is demonstrated in [7] that, for a single source, the MWF can be expressed as an MVDR beamformer followed by a single-channel Wiener filter. In addition, it is shown in [18] and [19] that many conventional multichannel filters such as MVDR, LCMV and MWF, are special cases of the variable span filters, which formulates the filter in terms of subspaces defined by the joint diagonalization of the speech and noise covariance matrices.

The temporal statistics of speech have also been exploited in conventional approaches including adaptive beamformers, post-filtering, GSC and MWF to derive optimal spatio-temporal filters under various criteria. However, these methods use the second-order-statistics (SOS) of the speech and noise to compute the statistical optimal filter, which inherently assumes the speech to be short-time stationary. Therefore, the short-time temporal evolution of speech is not considered. In fact, the speech signals in successive frames are correlated and are typically modelled as an auto-regressive (AR) process [20]. In the multichannel case, the inter-channel spatial correlation and inter-frame temporal correlation of speech can be jointly used to estimate the clean speech.

Manuscript received April 19, 2020; revised October 7, 2020; accepted November 13, 2020. Date of publication November 27, 2020; date of current version December 21, 2020. This work was supported by the Engineering and Physical Sciences Research Council projects EP/M026698/1 and EP/S035842/1. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zhijian Ou. *(Corresponding author: Wei Xue.)*

Wei Xue is with JD AI Research, Beijing , China, and also with the Imperial College London, London SW7 2AZ, U.K. (e-mail: xuewei.x@gmail.com).

Alastair H. Moore, Mike Brookes, and Patrick A. Naylor are with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: alastair.h.moore; mike.brookes@imperial.ac.uk; p.naylor@imperial.ac.uk).

Digital Object Identifier 10.1109/TASLP.2020.3040850

Many methods [21]–[39] have been proposed to incorporate the temporal evolution of speech into Kalman filtering (KF) for single-channel speech enhancement. Time-domain methods were developed in [21]–[25], and [22] demonstrated improved quality and intelligibility in both metric and listener-based evaluation, by using a codebook approach to estimate the AR coefficients of speech. Subband processing was used to reduce the order of AR models of speech [26]–[28], and the idea was extended to the short-time Fourier transform (STFT) domain processing [29] which models the temporal evolution of the complex STFT coefficients in each frequency bin. It was found that the phases in successive STFT frames are almost uncorrelated [29], thus in [30], the same authors chose to model only the evolution of STFT amplitudes in KF. Similarly, modulation-domain KF (MDKF) based methods have been proposed [31]–[39] in which the time-varying amplitude of each frequency bin is regarded as a modulation signal in its own right.

Inspired by these previous single-channel KF based speech enhancement methods, we proposed a modulation-domain multichannel Kalman filtering (MKF) speech enhancement method in [40], [41] that jointly exploits both the inter-channel spatial correlation and the inter-frame temporal correlation of speech. Based on the temporal evolution model of speech in the modulation domain, an STFT-domain linear prediction (LP) estimate is obtained by first performing LP in the modulation domain, and then inserting the phase from the MVDR beamformer output. Under the minimum mean squared error (MMSE) criterion, an optimal MKF gain is derived to combine the STFT-domain LP estimation and multichannel noisy observations for estimating the clean target signal. It is shown in [41] that the MKF becomes the MWF if the LP information is not incorporated, and by using the phase of the MVDR to approximate the phase of the clean speech, the MKF is equivalent to a concatenation of an MVDR beamformer and a single-channel MDKF.

An important problem of speech enhancement is to compromise between the noise reduction and the speech distortion. Although multichannel speech distortionless filters such as MVDR [2], [4], [6] and LCMV [5] beamformers can be derived, more aggressive noise reduction still gives rise to speech distortion in the output signal. However, the trade-off behaviour varies for different algorithms, and it is possible to obtain substantially improved noise reduction by accepting a limited amount of additional speech distortion [42], [43].

The requirements for noise reduction and speech distortion always vary between different applications and even between time-frequency (TF) bins for STFT-domain methods. For example, an algorithm might perform aggressive noise reduction in noise-dominated TF bins, but limit the speech distortion in other bins. Multichannel approaches which enable a flexible control of the trade-off behaviour in individual TF bins have been developed. The speech distortion weighted MWF (SDW-MWF) was proposed in [14], [44]; the optimal filter is obtained by minimizing a speech distortion index while requiring the noise reduction factor be below a specific threshold, which is related to a controlling parameter. A spherical harmonic-domain solution was further presented in [45]. Different variations of SDW-MWF have also been proposed to determine the controlling parameter in each TF bin according to voice activity detection (VAD) [46], conditional speech presence probability (SPP) [47], [48] and direct-to-reverberation ratio (DRR) [49]. In [18], [19], it was shown that flexible control of speech distortion and noise reduction can also be achieved within the variable span filtering framework of which the SDW-MWF is a special case.

It has been shown that, to achieve aggressive noise reduction in noise-dominated TF bins, the SDW-MWF tends to yield a zero-valued output signal [14], therefore, the speech component is also eliminated regardless of whether it is actually present, which leads to speech distortion. According to the analysis in [41], the MKF can be seen as integrating the temporal evolution of speech into the conventional MWF, and the optimal MKF gain weights between the LP estimation of the clean speech and the noisy observation to yield the output signal. Larger weight will be given to the LP estimation in noise-dominated TF bins, thus the speech component can be better preserved in the MKF output. If we design the trade-off filter based on MKF, the speech distortion can be better controlled when achieving noise reduction.

In this paper, a modulation-domain parametric MKF (PMKF) is proposed that extends the MKF. Based on the decomposition of the MKF cost function, a new cost function for PMKF is proposed, which uses a controlling parameter to weight the noise reduction and speech distortion related terms in the MKF cost function. An optimal PMKF gain is derived under the minimum mean squared error (MMSE) criterion. We analyse the performance of the proposed PMKF and show its relation to the speech distortion weighted multichannel Wiener filtering (SDW-MWF). It is shown that, by exploiting the speech evolution, the PMKF can always yield lower residual noise than the SDW-MWF when achieving the same amount of speech distortion. To evaluate the impact of the controlling parameter on the speech enhancement performance, we conduct experiments on a publicly available head-related impulse response (HRIR) database in different noisy and reverberant conditions, and present experimental results for both fixed and adaptive controlling parameters to demonstrate the effectiveness of the proposed method.

This paper is an extension of our previous work in [50]; it gives a more detailed derivation of the PMKF, shows explicitly how the controlling parameter affects the speech distortion and noise reduction, and conducts more comprehensive experimental evaluations. The remainder of the paper is organized as follows. In Section II, the signal model and assumptions are introduced. We review the previously proposed MKF in Section III, and derive the proposed PMKF in Section IV. The trade-off effect of PMKF on noise reduction and speech distortion is analysed in Section V and are compared with the SDW-MWF. The experimental results are shown in Section VI and we draw conclusions in Section VII.

II. Signal Model

We consider a noisy and reverberant environment with a single target speech source and an M-element microphone array. The complex STFT-domain $M \times 1$ noisy signal vector in the n-th frame and k-th frequency bin can be expressed as

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

where $\mathbf{y}(n,k) = [Y_1(n,k),\ Y_2(n,k),\ \ldots,\ Y_M(n,k)]^T$ is the noisy signal vector, in which $Y_m(n,k)$ is the STFT-domain noisy signal of the m-th microphone. The vectors $\mathbf{x}(n,k)$ and $\mathbf{v}(n,k)$ are defined similarly to denote the target reverberant speech and noise, respectively. If the RIR is longer than the analysis window, the late reverberation is treated as a component of $\mathbf{v}(n,k)$, and is assumed to be uncorrelated with direct-path signal and early reflections (see e.g. [51], [52]). We also assume that the speech and additive noise signals are uncorrelated.

The noisy signal vector, $\mathbf{y}(n,k)$, can be rewritten with respect to a reference signal $S(n,k)$, as

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

where $\mathbf{d}(k) = [D_1(k),\ D_2(k),\ \ldots,\ D_M(k)]^T$, and is the acoustic transfer function (ATF) vector if $S(n,k)$ is the speech source, or the relative transfer function (RTF) vector [53] with respect to the first channel if $S(n,k) = X_1(n,k)$. Because generally the RTF is shorter than the ATF when using a compact array, and the RTF can be estimated in noise by methods such as [54], [55], we take $S(n,k) = X_1(n,k)$ and denote $\mathbf{d}(k)$ as the RTF in this paper, then $D_m(k) = X_m(n,k)/X_1(n,k)$ for $m = 1, 2, \ldots, M$, and $D_1(k) = 1$. Similar to the beamforming, GSC and post-filtering methods, it is assumed here that the RTF vector is known or has been estimated. In this paper we only consider the noise reduction problem, and take $X_1(n,k)$ as the target. We do not explicitly perform dereverberation unless when the RIR is longer than the analysis window, $X_1(n,k)$ is the mixture of direct-path signal and early reverberation.

III. MKF for Speech Enhancement

We proposed a modulation-domain MKF in [40], [41] that exploits both the temporal evolution of speech and the spatial correlation between multiple microphones for speech enhancement. Modulation-domain processing treats the time-varying amplitude envelope in each frequency bin as a time-domain signal, and has been widely used for single-channel KF based speech enhancement [31]–[39] due both to its psychoacoustic and physiological significance [56], [57] and to the fact that the temporal correlation is mainly manifested in the magnitude spectrum [29].

Following the framework of conventional KF [58], the MKF first obtains an LP estimation of the hidden state, which represents the clean speech signal, and then updates the state by exploiting the multichannel noisy observation which contains the spatial information. It is not possible to use a conventional KF to estimate the state vector, because the temporal evolution and the spatial information are exploited in the modulation domain and the STFT domain respectively, and the mapping between these two domains is non-linear. An MKF which iteratively performs optimal LP estimation and state update in both domains was derived in [41], and its details will be briefly reviewed in this section, as it provides the foundation for the proposed PMKF given in Section IV.

A. State-Space Model

The MKF utilizes a measurement model and a P-order LP model to describe respectively the relationship between the target and observation, and the temporal speech evolution.

Based on the signal model in (2), the STFT-domain multichannel observation model can be defined as a function of the $P \times 1$ state vector $\mathbf{x}_1(n,k) = [X_1(n,k)\ X_1(n-1,k)\ \ldots\ X_1(n-P+1,k)]^T$:

$$\begin{aligned}\mathbf{y}(n,k) &= \mathbf{d}(k)X_1(n,k) + \mathbf{v}(n,k)\\ &= \mathbf{d}(k)\mathbf{u}^T\mathbf{x}_1(n,k) + \mathbf{v}(n,k)\\ &= \mathbf{Q}(k)\mathbf{x}_1(n,k) + \mathbf{v}(n,k),\end{aligned} \tag{3}$$

where $\mathbf{Q}(k) = \mathbf{d}(k)\mathbf{u}^T$ is an $M \times P$ measurement matrix, and $\mathbf{u} = [1\ 0\ \ldots\ 0]^T$ is a $P \times 1$ vector.

We temporarily neglect the spatial information and define the LP model in terms of the modulation-domain signal, $|X_1(n,k)|$, of the reference channel. The modulation-domain P-order LP model is formulated as:

$$\mathbf{a}_1(n,k) = \mathbf{B}(k)\mathbf{a}_1(n-1,k) + \mathbf{u}W(n,k), \tag{4}$$

where $\mathbf{a}_1(n,k) = [A_1(n,k),\ A_1(n-1,k),\ \ldots,\ A_1(n-P+1,k)]^T$ is the magnitude vector of the first channel with $A_1(n,k) = |X_1(n,k)|$. $\mathbf{B}(k)$ is a speech transition matrix defined as [41],

$$\mathbf{B}(k) = \begin{bmatrix} -b_{1,k} & -b_{2,k} & \ldots & -b_{p-1,k} & -b_{p,k}\\ 1 & 0 & \ldots & 0 & 0\\ 0 & 1 & \ldots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix}, \tag{5}$$

where $b_{p,k}$ for $p = 1, 2, \ldots, P$ are the LP coefficients [53] in the k-th frequency bin, and $W(n,k)$ is the LP residual with variance σ_W^2.

In practice, $\mathbf{B}(k)$ and σ_W^2 are unknown and are estimated via LP analysis of the modulation frames [31], [37]. We perform MWF pre-processing to obtain an output $Z_1(n,k)$. Then, using the autocorrelation method [59], the LP coefficients of the modulating signal in each frequency bin are estimated using the magnitude of $Z_1(n,k)$. The MWF pre-processing is realized as an MVDR beamformer followed by a single-channel Wiener post-filter [60].

The frequency index, k, will be omitted from the rest of the paper for clarity. We note that whereas $\mathbf{u}$ is a constant vector, both the RTF, $\mathbf{d}$, and the measurement matrix, $\mathbf{Q}$, are frequency-dependent.

B. MKF Solution

A block diagram of the MKF is shown in Fig. 1. Given the LP model, an *a priori* STFT-domain LP estimate $\hat{\mathbf{x}}_1(n|n-1)$ is obtained, by first performing LP in the modulation domain, and then transforming the modulation-domain LP estimation into the STFT domain after integrating the phase information $\hat{\mathbf{\Phi}}(n)$. By incorporating the multichannel noisy observations, the

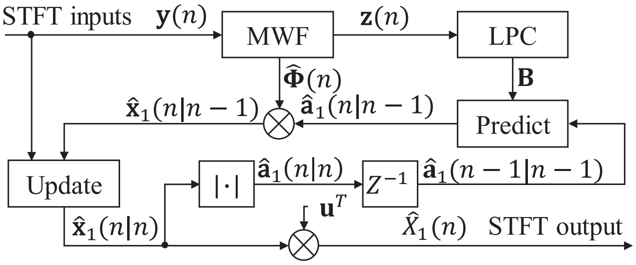

Fig. 1. MKF framework.

STFT-domain optimal MKF gain is derived under the MMSE criterion to update the STFT-domain state vector.

1) STFT-Domain Lp: Given the STFT-domain MMSE estimation of the clean speech in the previous frame, $\hat{\mathbf{x}}_1(n-1|n-1)$, the "predict" block estimates the amplitude of the state vector in the current frame according to (4) as:

$$\hat{\mathbf{a}}_1(n|n-1) = \mathbf{B}\hat{\mathbf{a}}_1(n-1|n-1), \tag{6}$$

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

We next obtain the STFT-domain LP estimation by imposing the phase of the MWF pre-processed output, $\mathbf{z}_1(n)$, as

$$\hat{\mathbf{x}}_1(n|n-1) = \hat{\mathbf{\Phi}}(n)\hat{\mathbf{a}}_1(n|n-1), \tag{7}$$

where $\hat{\mathbf{\Phi}}(n)$ is a diagonal phase matrix containing the complex exponential phase of $\mathbf{z}_1(n)$.

2) STFT-Domain Update: In the "update" block of Fig. 1, the STFT-domain state vector is updated by linearly combining the estimates from the STFT-domain LP and from the multichannel noisy observations as:

$$\mathbf{x}_1(n|n) = \mathbf{x}_1(n|n-1) + \mathbf{G}(n)[\mathbf{y}(n) - \mathbf{Q}\mathbf{x}_1(n|n-1)], \tag{8}$$

where $\mathbf{G}(n)$ is the MKFs gain matrix with dimension $P \times M$.

To determine the optimal MKF gain matrix, we define the error signal vector $\mathbf{e}_1(n|n)$ between the updated state vector $\mathbf{x}_1(n|n)$ and clean speech vector $\mathbf{x}_1(n)$. Based on (8), we have

$$\begin{aligned}
&\mathbf{e}_1(n|n)\\
&= \mathbf{x}_1(n|n) - \mathbf{x}_1(n)\\
&= \mathbf{x}_1(n|n-1) - \mathbf{x}_1(n) + \mathbf{G}(n)[\mathbf{y}(n) - \mathbf{Q}\mathbf{x}_1(n|n-1)]\\
&= \mathbf{e}_1(n|n-1) + \mathbf{G}(n)[\mathbf{v}(n) - \mathbf{Q}\mathbf{e}_1(n|n-1)]\\
&= [\mathbf{I} - \mathbf{G}(n)\mathbf{Q}]\mathbf{e}_1(n|n-1) + \mathbf{G}(n)\mathbf{v}(n),
\end{aligned} \tag{9}$$

where $\mathbf{e}_1(n|n-1)$ is a STFT-domain LP estimation error vector given by

$$\mathbf{e}_1(n|n-1) = \mathbf{x}_1(n|n-1) - \mathbf{x}_1(n). \tag{10}$$

By minimizing an MMSE based cost function

$$J_{\text{MKF},\mathbf{G}(n)} = \text{tr}[\mathbf{R}_{ee}(n|n)], \tag{11}$$

where $\mathbf{R}_{ee}(n|n) = \mathbb{E}\{\mathbf{e}_1(n|n)\mathbf{e}_1^H(n|n)\}$ and $\mathbb{E}\{\cdot\}$ is the expectation operator, we obtain the optimal MKF gain $\hat{\mathbf{G}}_{\text{MKF}}(n)$:

$$\begin{aligned}
&\hat{\mathbf{G}}_{\text{MKF}}(n)\\
&= \mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H[\mathbf{Q}\mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H + \mathbf{R}_{vv}(n)]^{-1},
\end{aligned} \tag{12}$$

where $\mathbf{R}_{ee}(n|n-1) = \mathbb{E}\{\mathbf{e}_1(n|n-1)\mathbf{e}_1^H(n|n-1)\}$ is the covariance matrix of the STFT-domain LP estimation error, and $\mathbf{R}_{vv}(n) = \mathbb{E}\{\mathbf{v}(n)\mathbf{v}^H(n)\}$ is the multichannel noise covariance matrix. The $\mathbf{R}_{vv}(n)$ can be estimated by methods described in, e.g., [49], [61], [62], and the estimation and updating of $\mathbf{R}_{ee}(n|n-1)$ are given the Section IV-C in [40].

After updating the state vector in the STFT domain, the clean speech signal of the reference channel is finally estimated as $\hat{X}_1(n) = \mathbf{u}^T\mathbf{x}_1(n|n)$.

IV. Proposed PMKF for Speech Enhancement

A theoretical analysis of the MKF was included in [41] where it was shown that the MKF can be viewed as incorporating knowledge of speech evolution into MWF. The MKF gain in (12) is controlled by the variance of the STFT-domain LP error and the noise level, and from (8) we can deduce that the output signal slides between the LP estimation and the estimation from multichannel noisy observations. The output will approximate the LP estimate when the noise level is high, and approximate the observation when the noise level is low, in which case the multichannel observations are more reliable. By incorporating the temporal evolution of speech, the noise reduction behaviour of MKF is different from that of the MWF, which computes the Wiener gain solely based on the relative level of speech and noise and, in high-noise scenarios, gives a near-zero gain such that all signal components are eliminated. Compared with MWF, more speech information can be preserved by MKF, which yields less speech distortion.

The trade-off filter for multichannel speech enhancement is useful since the requirement for speech distortion and noise reduction varies for different applications, and the trade-off behaviour in each TF bin can be flexibly controlled. The trade-off filter can be designed, for instance, by following the principles of SDW-MWF, or more generally, variable span filtering [18], [19]. Motivated by the comparison between MKF and MWF, a PMKF is proposed in this section, which uses a controlling parameter to select a trade-off between speech distortion and noise reduction.

A. Cost Function of PMKF

The proposed PMKF utilizes the same state-space model as the MKF described in Section III-A, which defines the state vector as the STFT-domain clean signal vector of the reference channel. It models the temporal evolution of speech in the modulation domain, but uses an STFT-domain measurement model to incorporate spatial information.

The PKMF uses the same prediction and equations as the MKF given in (6), (7) but in the update step, (8), it uses a different gain matrix, $\mathbf{G}(n)$, which is derived below.

By substituting (1) into (8), the output signal vector, $\mathbf{x}_1(n|n)$, can be expressed as

$$\begin{aligned}\mathbf{x}_1(n|n) &= [\mathbf{I} - \mathbf{G}(n)\mathbf{Q}]\mathbf{x}_1(n|n-1) + \mathbf{G}(n)[\mathbf{x}(n) + \mathbf{v}(n)] \\ &= [\mathbf{I} - \mathbf{G}(n)\mathbf{Q}]\mathbf{x}_1(n|n-1) + \mathbf{G}(n)\mathbf{Q}\mathbf{x}_1(n) \\ &\quad + \mathbf{G}(n)\mathbf{v}(n), \qquad (13)\end{aligned}$$

in which the output vector has been decomposed into a speech-related component, $[\mathbf{I} - \mathbf{G}(n)\mathbf{Q}]\mathbf{x}_1(n|n-1) + \mathbf{G}(n)\mathbf{Q}\mathbf{x}_1(n)$, and a residual noise component, $\mathbf{G}(n)\mathbf{v}(n)$. The speech-related component is a linear combination of the STFT-domain LP estimate and clean speech vector of the reference channel, with weights that are defined by the matrix $\mathbf{G}(n)\mathbf{Q}$.

We now define a speech distortion vector $\boldsymbol{\nu}_{\text{sd},\mathbf{G}(n)}$, to be the difference between the speech-related component and the target $\mathbf{x}_1(n)$, which is the clean speech in the first channel:

$$\begin{aligned}&\boldsymbol{\nu}_{\text{sd},\mathbf{G}(n)} \\ &= \{[\mathbf{I} - \mathbf{G}(n)\mathbf{Q}]\mathbf{x}_1(n|n-1) + \mathbf{G}(n)\mathbf{Q}\mathbf{x}_1(n)\} - \mathbf{x}_1(n) \\ &= [\mathbf{I} - \mathbf{G}(n)\mathbf{Q}][\mathbf{x}_1(n|n-1) - \mathbf{x}_1(n)] \\ &= [\mathbf{I} - \mathbf{G}(n)\mathbf{Q}]\mathbf{e}_1(n|n-1), \qquad (14)\end{aligned}$$

which is actually the first term of the STFT-domain error signal vector $\mathbf{e}_1(n|n)$ of the MKF in (9).

Since the $\mathbf{e}_1(n|n-1)$ is calculated solely from the speech signal, based on the assumption that the speech and noise are uncorrelated, we use (9) to rewrite the MMSE based cost function for MKF in (11) as

$$\begin{aligned}J_{\text{MKF},\mathbf{G}(n)} &= \underbrace{\text{tr}\{[\mathbf{I} - \mathbf{G}(n)\mathbf{Q}]\mathbf{R}_{ee}(n|n-1)[\mathbf{I} - \mathbf{G}(n)\mathbf{Q}]^H\}}_{J_x[\mathbf{G}(n)]} \\ &\quad + \underbrace{\text{tr}\{\mathbf{G}(n)\mathbf{R}_{vv}(n)\mathbf{G}^H(n)\}}_{J_v[\mathbf{G}(n)]}, \qquad (15)\end{aligned}$$

where $J_x[\mathbf{G}(n)]$ and $J_v[\mathbf{G}(n)]$ represent the speech distortion and the residual noise in the MKF output, respectively.

In order to trade off between the speech distortion and noise reduction, now we propose a new MMSE based cost function for the PMKF, as a weighted combination of $J_x[\mathbf{G}(n)]$ and $J_v[\mathbf{G}(n)]$,

$$J_{\text{PMKF},\mathbf{G}(n)} = J_x[\mathbf{G}(n)] + \lambda J_v[\mathbf{G}(n)], \qquad (16)$$

where $\lambda > 0$ is the controlling parameter of PMKF. Comparing with (15), it can be seen that when $\lambda = 1$, the cost function of PMKF is identical to the MKF. If $\lambda > 1$, more emphasis will be given to noise reduction, and if $\lambda < 1$, more emphasis will be given to controlling the speech distortion.

B. PMKF Solution

The optimal PMKF gain matrix is obtained by minimizing the PMKF cost function $J_{\text{PMKF},\mathbf{G}(n)}$ with respect to $\mathbf{G}(n)$ to obtain [63]

$$\begin{aligned}&\hat{\mathbf{G}}_{\text{PMKF}}(n) = \arg\min_{\mathbf{G}(n)} J_{\text{PMKF},\mathbf{G}(n)} \\ &= \mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H[\mathbf{Q}\mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H + \lambda\mathbf{R}_{vv}(n)]^{-1}. \qquad (17)\end{aligned}$$

It can be seen that the PMKF gain matrix has a similar form to the MKF gain matrix in (12), but now incorporates a parameter, λ, to control the trade-off between speech distortion and noise reduction.

From the identity $\mathbf{Q} = \mathbf{d}\mathbf{u}^T$ in (3), we can write,

$$\mathbf{Q}\mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H = \sigma_e^2(n|n-1)\mathbf{d}\mathbf{d}^H, \qquad (18)$$

where

$$\begin{aligned}\sigma_e^2(n|n-1) &= \mathbb{E}\{E_1(n|n-1)E_1^*(n|n-1)\} \\ &= \mathbf{u}^T\mathbf{R}_{ee}(n|n-1)\mathbf{u} \qquad (19)\end{aligned}$$

is the first diagonal element of $\mathbf{R}_{ee}(n|n-1)$, and

$$E_1(n|n-1) = \mathbf{u}^T\mathbf{e}_1(n|n-1) \qquad (20)$$

is the first element of the STFT-domain LP error vector $\mathbf{e}_1(n|n-1)$. Thus, $\mathbf{Q}\mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H$ is a rank-one matrix, and the matrix that is inverted in (17) will be close to singular when λ is small or the noise level is low. The same numerical problem also occurs when computing the MKF gain matrix in (12).

To avoid this numerical problem, in the implementation of all algorithms, we compute the inverse of a matrix $\mathbf{P} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^H$ based on the singular value decomposition, as $\mathbf{P}^{-1} = \mathbf{V}\bar{\boldsymbol{\Lambda}}\mathbf{U}^H$. Here $\bar{\boldsymbol{\Lambda}}$ is a diagonal matrix whose i-th diagonal element is $1/\boldsymbol{\Lambda}_{i,i}$ if $|\boldsymbol{\Lambda}_{i,i}| > \max_i\{|\boldsymbol{\Lambda}_{i,i}|\} \times \zeta$, and is 0 otherwise. The $\boldsymbol{\Lambda}_{i,i}$ is the i-th diagonal element of $\boldsymbol{\Lambda}$, and in practice, ζ can be chosen according to the uncertainty of the RTF vector $\mathbf{d}$.

The optimal PMKF gain matrix $\hat{\mathbf{G}}_{\text{PMKF}}(n)$ is substituted into (8) to update the state vector, and the covariance matrix of the STFT-domain LP estimation error $\mathbf{R}_{ee}(n|n-1)$ is also updated using $\hat{\mathbf{G}}_{\text{PMKF}}(n)$ based on Section IV. C of [41]. Finally, as in [41], the clean speech of the reference channel is estimated as $\hat{X}_1(n) = \mathbf{u}^T\mathbf{x}_1(n|n)$.

V. Performance Analysis

The performance of the proposed PMKF will be analysed in this section. First we show how the controlling parameter λ in PMKF affects the speech distortion and noise reduction of the PMKF output. Then we compare the proposed PMKF with the SDW-MWF [14], [44], which similarly uses a parameter to control the trade-off between speech distortion and noise reduction, but is derived in a Wiener filtering framework.

A. Effect of Controlling Parameter on Speech Distortion

Since the output of the PMKF is calculated by $\hat{X}_1(n) = \mathbf{u}^T\mathbf{x}_1(n|n)$, and the target signal is $X_1(n) = \mathbf{u}^T\mathbf{x}_1(n)$, the speech distortion signal of the PMKF output, $\epsilon_{\text{sd,PMKF}}(n)$, is defined as the first element of the speech distortion vector in (14):

$$\begin{aligned}\epsilon_{\text{sd,PMKF}}(n) &= \mathbf{u}^T\boldsymbol{\nu}_{\text{sd}}[\hat{\mathbf{G}}_{\text{PMKF}}(n)] \\ &= \mathbf{u}^T[\mathbf{I} - \hat{\mathbf{G}}_{\text{PMKF}}(n)\mathbf{Q}]\mathbf{e}_1(n|n-1) \\ &= E_1(n|n-1) - \mathbf{u}^T\hat{\mathbf{G}}_{\text{PMKF}}(n)\mathbf{Q}\mathbf{e}_1(n|n-1), \qquad (21)\end{aligned}$$

where $E_1(n|n-1)$ is defined in (20).

We now consider the second term in the right side of (21). Substituting for $\mathbf{G}_{\text{PMKF}}(n)$ using (17), (18) and the identity $\mathbf{u}^T\mathbf{d} = 1$, we obtain

$$\begin{aligned}
&\mathbf{u}^T\hat{\mathbf{G}}_{\text{PMKF}}(n)\mathbf{Q}\mathbf{e}_1(n|n-1)\\
&= (\mathbf{u}^T\mathbf{d})\mathbf{u}^T\mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H\\
&\quad\times[\sigma_e^2(n|n-1)\mathbf{d}\mathbf{d}^H + \lambda\mathbf{R}_{vv}(n)]^{-1}\mathbf{Q}\mathbf{e}_1(n|n-1)\\
&= \mathbf{u}^T\sigma_e^2(n|n-1)\mathbf{d}\mathbf{d}^H\\
&\quad\times[\sigma_e^2(n|n-1)\mathbf{d}\mathbf{d}^H + \lambda\mathbf{R}_{vv}(n)]^{-1}\mathbf{d}\mathbf{u}^T\mathbf{e}_1(n|n-1)\\
&= \sigma_e^2(n|n-1)\mathbf{d}^H[\sigma_e^2(n|n-1)\mathbf{d}\mathbf{d}^H\\
&\quad+\lambda\mathbf{R}_{vv}(n)]^{-1}\mathbf{d}E_1(n|n-1). \qquad (22)
\end{aligned}$$

According to the Sherman-Morrison-Woodbury formula [64],

$$\begin{aligned}
&[\sigma_e^2(n|n-1)\mathbf{d}\mathbf{d}^H + \lambda\mathbf{R}_{vv}(n)]^{-1}\\
&= \lambda^{-1}\mathbf{R}_{vv}^{-1}(n) - \frac{\lambda^{-2}\sigma_e^2(n|n-1)\mathbf{R}_{vv}^{-1}(n)\mathbf{d}\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)}{1+\lambda^{-1}\sigma_e^2(n|n-1)\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbf{d}}, \qquad (23)
\end{aligned}$$

then

$$\begin{aligned}
&\mathbf{d}^H[\sigma_e^2(n|n-1)\mathbf{d}\mathbf{d}^H + \lambda\mathbf{R}_{vv}(n)]^{-1}\mathbf{d}\\
&= \lambda^{-1}\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbf{d}\\
&\quad - \frac{\lambda^{-2}\sigma_e^2(n|n-1)\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbf{d}\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbf{d}}{1+\lambda^{-1}\sigma_e^2(n|n-1)\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbf{d}}\\
&= \frac{\lambda^{-1}\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbf{d}}{1+\lambda^{-1}\sigma_e^2(n|n-1)\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbf{d}}. \qquad (24)
\end{aligned}$$

Note that

$$\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbf{d} = \sigma_{V_{o,\text{MVDR}}}^{-2}(n), \qquad (25)$$

where $\sigma_{V_{o,\text{MVDR}}}^2(n)$ is the variance of the residual noise in the MVDR beamformer output [see [41] Eq. (43)]. Thus (24) becomes

$$\begin{aligned}
&\mathbf{d}^H[\sigma_e^2(n|n-1)\mathbf{d}\mathbf{d}^H + \lambda\mathbf{R}_{vv}(n)]^{-1}\mathbf{d}\\
&= \frac{\lambda^{-1}\sigma_{V_{o,\text{MVDR}}}^{-2}(n)}{1+\lambda^{-1}\sigma_{V_{o,\text{MVDR}}}^{-2}(n)\sigma_e^2(n|n-1)}. \qquad (26)
\end{aligned}$$

Substituting (26) into (22), the speech distortion signal in (21) is now expressed as

$$\begin{aligned}
&\epsilon_{\text{sd,PMKF}}(n)\\
&= E_1(n|n-1) - \frac{\lambda^{-1}\sigma_{V_{o,\text{MVDR}}}^{-2}(n)\sigma_e^2(n|n-1)E_1(n|n-1)}{1+\lambda^{-1}\sigma_{V_{o,\text{MVDR}}}^{-2}(n)\sigma_e^2(n|n-1)}\\
&= \frac{E_1(n|n-1)}{1+\lambda^{-1}\sigma_{V_{o,\text{MVDR}}}^{-2}(n)\sigma_e^2(n|n-1)}. \qquad (27)
\end{aligned}$$

The variance of the speech distortion signal, $\sigma_{\text{sd,PMKF}}^2(n)$, is given by

$$\begin{aligned}
&\sigma_{\text{sd,PMKF}}^2(n)\\
&= \mathbb{E}\{\epsilon_{\text{sd,PMKF}}(n)\epsilon_{\text{sd,PMKF}}^*(n)\}\\
&= \frac{\sigma_e^2(n|n-1)}{[1+\lambda^{-1}\sigma_{V_{o,\text{MVDR}}}^{-2}(n)\sigma_e^2(n|n-1)]^2}\\
&= \left[\frac{\lambda\sigma_{V_{o,\text{MVDR}}}^2(n)}{\lambda\sigma_{V_{o,\text{MVDR}}}^2(n)+\sigma_e^2(n|n-1)}\right]^2\sigma_e^2(n|n-1)\\
&= [1-g_{\text{PMKF}}(n)]^2\sigma_e^2(n|n-1), \qquad (28)
\end{aligned}$$

where

$$g_{\text{PMKF}}(n) = \frac{\sigma_e^2(n|n-1)}{\lambda\sigma_{V_{o,\text{MVDR}}}^2(n)+\sigma_e^2(n|n-1)} \qquad (29)$$

is a single-channel Wiener-type gain which is defined according to the noise level of the MVDR output and the variance of STFT-domain LP error, and is also affected by the controlling parameter λ of the PMKF.

In (28) we express the speech distortion in the PMKF output as a monotonic function of the controlling parameter λ. It can be seen that $\sigma_{\text{sd,PMKF}}^2(n) \rightarrow 0$ when $\lambda \rightarrow 0$, indicating that there is no distortion in the output signal. This is consistent with the cost function design in (16) that decreasing the controlling parameter gives more emphasis on limiting the speech distortion. When $\lambda \rightarrow +\infty$, we have $\sigma_{\text{sd,PMKF}}^2(n) \rightarrow \sigma_e^2(n|n-1)$, in this case, the speech distortion is only caused by STFT-domain LP error.

Since $\lambda \geq 0$, it follows that $0 < g_{\text{PMKF}}(n) \leq 1$, and the speech distortion variance in (28) is always smaller than $\sigma_e^2(n|n-1)$. Therefore, by incorporating the noisy observations, the speech distortion is always reduced in the update step of PMKF. We can further deduce that, if the STFT-domain clean speech signal can be precisely estimated in the LP step, which makes $\sigma_e^2(n|n-1) = 0$, the PMKF will yield an output signal without any speech distortion.

It can be also seen from (28) that the noise level affects the distortion of the PMKF output. In high-noise cases, according to (26), the large elements of $\mathbf{R}_{vv}$ leads to a large value of $\sigma_{V_{o,\text{MVDR}}}^2(n)$, which further decreases $g_{\text{PMKF}}(n)$, and finally yields a large speech distortion in (28). However, by introducing the parameter λ, the speech distortion behaviour of PMKF can now be flexibly controlled.

B. Effect of Controlling Parameter on Noise Reduction

To examine the noise reduction performance, we calculate the residual noise in the PMKF output. From (13), with the optimal PMKF gain matrix, the residual noise in the PMKF output is expressed as

$$V_{o,\text{PMKF}}(n) = \mathbf{u}^T\hat{\mathbf{G}}_{\text{PMKF}}(n)\mathbf{v}(n). \qquad (30)$$

Similar to the derivation in (22), applying (23) and (25), we have

$$\begin{aligned}
&V_{o,\text{PMKF}}(n)\\
&= \mathbf{u}^T\mathbf{Q}\mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H\\
&\quad\times[\sigma_e^2(n|n-1)\mathbf{d}\mathbf{d}^H + \lambda\mathbf{R}_{vv}(n)]^{-1}\mathbf{v}(n)\\
&= \mathbf{u}^T\sigma_e^2(n|n-1)\mathbf{d}\mathbf{d}^H[\lambda^{-1}\mathbf{R}_{vv}^{-1}(n)
\end{aligned}$$

$$
\begin{aligned}
&\quad - \frac{\lambda^{-2}\sigma_e^2(n|n-1)\mathbf{R}_{vv}^{-1}(n)\mathbf{d}\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)}{1+\lambda^{-1}\sigma_e^2(n|n-1)\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbf{d}}]\mathbf{v}(n) \\
&= \lambda^{-1}\sigma_e^2(n|n-1)\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbf{v}(n) \\
&\quad - \frac{\lambda^{-2}\sigma_{V_{o,\text{MVDR}}}^{-2}(n)\sigma_e^4(n|n-1)}{1+\lambda^{-1}\sigma_{V_{o,\text{MVDR}}}^{-2}(n)\sigma_e^2(n|n-1)}\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbf{v}(n) \\
&= \frac{\lambda^{-1}\sigma_e^2(n|n-1)}{1+\lambda^{-1}\sigma_{V_{o,\text{MVDR}}}^{-2}(n)\sigma_e^2(n|n-1)}\mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbf{v}(n). \qquad (31)
\end{aligned}
$$

Then the variance of the residual noise $V_{o,\text{PMKF}}(n)$ is computed as

$$
\begin{aligned}
&\sigma_{V_{o,\text{PMKF}}}^2(n) \\
&= \mathbb{E}\{V_{o,\text{PMKF}}(n)V_{o,\text{PMKF}}^*(n)\} \\
&= \left[\frac{\lambda^{-1}\sigma_e^2(n|n-1)}{1+\lambda^{-1}\sigma_{V_{o,\text{MVDR}}}^{-2}(n)\sigma_e^2(n|n-1)}\right]^2 \\
&\quad \times \mathbf{d}^H\mathbf{R}_{vv}^{-1}(n)\mathbb{E}\{\mathbf{v}(n)\mathbf{v}^H(n)\}\mathbf{R}_{vv}^{-1}(n)\mathbf{d} \\
&= \left[\frac{\lambda^{-1}\sigma_e^2(n|n-1)}{1+\lambda^{-1}\sigma_{V_{o,\text{MVDR}}}^{-2}(n)\sigma_e^2(n|n-1)}\right]^2 \sigma_{V_{o,\text{MVDR}}}^{-2}(n) \\
&= \left[\frac{\sigma_e^2(n|n-1)}{\lambda\sigma_{V_{o,\text{MVDR}}}^2(n)+\sigma_e^2(n|n-1)}\right]^2 \sigma_{V_{o,\text{MVDR}}}^2(n) \\
&= g_{\text{PMKF}}^2(n)\sigma_{V_{o,\text{MVDR}}}^2(n). \qquad (32)
\end{aligned}
$$

The result in (32) reveals the relationship between the noise reduction performance between the PMKF and the MVDR beamformer. Since $g_{\text{PMKF}}(n) < 1$, in PMKF, the residual noise is further suppressed by applying a gain $g_{\text{PMKF}}(n)$ to the MVDR output. When $\lambda \to +\infty$, which corresponds to aggressive noise reduction, $g_{\text{PMKF}}(n) \to 0$, and the variance of the residual noise decreases to 0, indicating all noise components are suppressed. When decreasing λ to 0, $g_{\text{PMKF}}(n)$ gradually increases to 1, which attenuates less noise and finally yields the MVDR output.

According to (29) we further define $\tilde{\lambda}_{\text{PMKF}}(n) = \frac{\sigma_e^2(n|n-1)}{V_{o,\text{MVDR}}(n)}$, then $g_{\text{PMKF}}(n)$ becomes

$$
g_{\text{PMKF}}(n) = \frac{1}{\lambda/\tilde{\lambda}_{\text{PMKF}}(n)+1}, \qquad (33)
$$

and the speech distortion and noise reduction performances expressed in (28) and (32) are rewritten as

$$
\frac{\sigma_{\text{sd,PMKF}}^2(n)}{\sigma_e^2(n|n-1)} = \left(\frac{\lambda/\tilde{\lambda}_{\text{PMKF}}(n)}{\lambda/\tilde{\lambda}_{\text{PMKF}}(n)+1}\right)^2, \qquad (34)
$$

$$
\frac{\sigma_{V_{o,\text{PMKF}}}^2(n)}{\sigma_{V_{o,\text{MVDR}}}^2(n)} = \left(\frac{1}{\lambda/\tilde{\lambda}_{\text{PMKF}}(n)+1}\right)^2. \qquad (35)
$$

Based on (34) and (35), the speech distortion and noise reduction performances of PMKF are illustrated in Fig. 2. When increasing the controlling parameter λ, the PMKF yields less residual noise at the expense of more speech distortion. The variances of speech distortion and noise residual are upper-bounded by $\sigma_e^2(n|n-1)$ and $\sigma_{V_{o,\text{MVDR}}}^2(n)$, respectively.

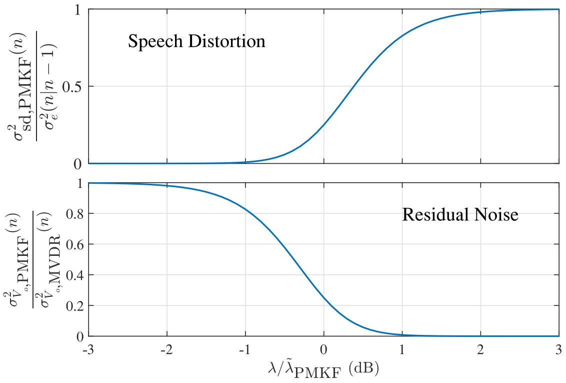

Fig. 2. Illustration of the theoretical trade-off behaviour of PMKF from (34) and (35).

C. Comparison With SDW-MWF

1) Relationship to SDW-MWF: We now first consider the relationship between the PMKF and the SDW-MWF presented in [14], [44]. The SDW-MWF is expressed as

$$
\mathbf{h}_{\text{SDW-MWF}}(n) = [\mathbf{R}_{xx}(n) + \lambda\mathbf{R}_{vv}(n)]^{-1}\mathbf{R}_{xx}(n)\mathbf{u}, \qquad (36)
$$

where $\mathbf{R}_{xx}(n) = \mathbb{E}\{\mathbf{x}(n)\mathbf{x}^H(n)\}$ is the speech covariance matrix, and λ controls the trade-off between speech distortion and noise suppression. With $\mathbf{h}_{\text{SDW-MWF}}(n)$, the clean speech of the reference channel is estimated as $\hat{X}_1(n) = \mathbf{h}_{\text{SDW-MWF}}^H(n)\mathbf{y}(n)$.

We first show that SDW-MWF can be regarded as a special case of the PMKF. If the LP information is excluded from the PMKF by setting the STFT-domain LP estimate $\mathbf{x}_1(n|n-1) \equiv \mathbf{0}$, then $\mathbf{Q}\mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H$ in (17) becomes the speech covariance matrix $\mathbf{R}_{xx}(n) = \mathbb{E}\{\mathbf{x}(n)\mathbf{x}^H(n)\}$ [41]. The optimal PMKF gain matrix is substituted into (8) in the update step. Since $\mathbf{u}^T\mathbf{d} = 1$, by setting the STFT-domain LP estimate to zero, the output signal of PMKF is derived as

$$
\begin{aligned}
\hat{X}_1(n) &= \mathbf{u}^T\hat{\mathbf{G}}_{\text{PMKF}}(n)\mathbf{y}(n) \\
&= (\mathbf{u}^T\mathbf{d})\mathbf{u}^T\mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H[\mathbf{R}_{xx}(n)+\lambda\mathbf{R}_{vv}(n)]^{-1}\mathbf{y}(n) \\
&= \mathbf{u}^T\mathbf{Q}\mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H[\mathbf{R}_{xx}(n)+\lambda\mathbf{R}_{vv}(n)]^{-1}\mathbf{y}(n) \\
&= \mathbf{u}^T\mathbf{R}_{xx}(n)[\mathbf{R}_{xx}(n)+\lambda\mathbf{R}_{vv}(n)]^{-1}\mathbf{y}(n) \\
&= \mathbf{h}_{\text{SDW-MWF}}^H\mathbf{y}(n). \qquad (37)
\end{aligned}
$$

Thus, as with the relationship between MKF and MWF discussed in [41], the proposed PMKF can be seen as incorporating the speech evolution over time into SDW-MWF.

2) Trade-Off Performance of SDW-MWF: Since SDW-MWF is a special case of the PMKF, (28) and (32) can straightforwardly be used to derive the speech distortion and noise reduction measures of the SDW-MWF. If the LP estimation is discarded, with the definition of $\mathbf{e}_1(n|n-1)$ in (10), we can replace $\sigma_e^2(n|n-1)$ by $\sigma_{X_1}^2(n)$, where $\sigma_{X_1}^2(n) = \mathbb{E}\{X_1(n)X_1^*(n)\}$.

According to (28), the variance of speech distortion in the SDW-MWF output is derived as

$$
\sigma_{\text{sd,SDW-MWF}}^2(n) = [1 - g_{\text{SDW-MWF}}(n)]^2\sigma_{X_1}^2(n), \qquad (38)
$$

where

$$g_{\text{SDW-MWF}}(n) = \frac{\sigma^2_{X_1}(n)}{\lambda\sigma^2_{V_{o,\text{MVDR}}}(n) + \sigma^2_{X_1}(n)}. \tag{39}$$

With (32), the variance of the residual noise in the SDW-MWF output is given by

$$\sigma^2_{V_{o,\text{SDW-MWF}}}(n) = g^2_{\text{SDW-MWF}}(n)\sigma^2_{V_{o,\text{MVDR}}}(n). \tag{40}$$

The effect of the controlling parameter λ on the trade-off behaviour of SDW-MWF can be analysed from (38) to (40). It can be seen that increasing λ decreases $g_{\text{SDW-MWF}}(n)$, and further leads to higher speech distortion and lower residual noise. In the extreme cases, when $\lambda \to 0$, similar to the PMKF, there is no speech distortion in the output signal, and the variance of the residual noise equals to that of the MVDR output. However, when $\lambda \to +\infty$, $g_{\text{SDW-MWF}}(n)$ approaches 0; in this case, although there is no residual noise in the SDW-MWF output, the speech component is also 0; in eliminated, it makes the variance of speech distortion $\sigma^2_{\text{sd,SDW-MWF}}(n)$ equal to $\sigma^2_{X_1}(n)$. This is different from the PMKF in which the speech distortion limit equals the STFT-domain LP estimation error. The trade-off behaviour of SDW-MWF is thus similar to that of PMKF in Fig. 2, except that the y-axes for speech distortion and residual noise variance are $\sigma^2_{\text{sd,SDW-MWF}}(n)/\sigma^2_{X_1}(n)$ and $\sigma^2_{V_{o,\text{SDW-MWF}}}(n)/\sigma^2_{V_{o,\text{MVDR}}}(n)$, respectively, and the x-axis becomes $\lambda/\tilde{\lambda}_{\text{SDW-MWF}}(n)$ where $\tilde{\lambda}_{\text{SDW-MWF}}(n) = \sigma^2_{X_1}(n)/V_{o,\text{MVDR}}(n)$.

We now consider the effect of RTF error on PMKF and SDW-MWF. From (40) it follows that, under the single-source assumption, the $\mathbf{R}_{xx}(n)$ in (36) can be rewritten as $\sigma^2_{X_1}(n)\mathbf{d}\mathbf{d}^H$, and the SDW-MWF can actually be factorized into an MVDR beamformer and a single-channel post-filter. The post-filter gain is $g_{\text{SDW-MWF}}(n)$ from (39). The explicit proof can be straightforwardly derived from Chapter 3.2.2 of [65] and is omitted here. Similarly, following Section V in [41], the proposed PMKF can also be factorized and implemented as a concatenation of an MVDR beamformer and a single-channel modulation-domain Kalman-type filter, whose trade-off behaviour is controlled by the parameter λ of PMKF. As a result, the RTF error mainly affects the MVDR beamforming stages in both methods; this has been analysed in [66].

3) Trade-Off Performance Comparison: The PMKF exploits the temporal evolution of speech to perform LP. From the above analysis, as long as the STFT-domain LP of PMKF provides a better estimate of the clean speech than a zero-valued signal, we have $\sigma^2_e(n|n-1) < \sigma^2_{X_1}(n)$, and the upper bound for the variance of speech distortion is smaller than that of SDW-MWF. Therefore, by incorporating the LP information, the PMKF has the potential to achieve lower speech distortion than SDW-MWF.

To analyse further the performance difference between PMKF and SDW-MWF, we compare the residual noise level in the output signals of the PMKF and SDW-MWF when the speech distortion is fixed at $K\sigma^2_{X_1}(n)$ for some K.

We assume that $\sigma^2_e(n|n-1) < \sigma^2_{X_1}(n)$, and express $\sigma^2_e(n|n-1)$ as

$$\sigma^2_e(n|n-1) = \rho\sigma^2_{X_1}(n), \tag{41}$$

where $0 < \rho < 1$. Let $K\sigma^2_{X_1}(n)$ denote the target speech distortion variance for both PMKF and SDW-MWF. From (28), (29), and (41), for the PMKF, we have

$$\left[1 - \frac{\rho\sigma^2_{X_1}(n)}{\lambda_{\text{PMKF}}\sigma^2_{V_{o,\text{MVDR}}}(n) + \rho\sigma^2_{X_1}(n)}\right]^2 \rho\sigma^2_{X_1}(n) = K\sigma^2_{X_1}(n), \tag{42}$$

where λ_{PMKF} is the controlling parameter of PMKF when $\sigma^2_{\text{sd,PMKF}}(n) = K\sigma^2_{X_1}(n)$, and is computed as

$$\lambda_{\text{PMKF}} = \frac{\sqrt{K}\rho\sigma^2_{X_1}(n)}{(\sqrt{\rho} - \sqrt{K})\sigma^2_{V_{o,\text{MVDR}}}(n)}. \tag{43}$$

Substituting (43) into (29) and (32), the variance of the residual noise in PMKF can be finally expressed as

$$\sigma^2_{V_{o,\text{PMKF}}}(n) = \left(1 - \sqrt{K/\rho}\right)^2 \sigma^2_{V_{o,\text{MVDR}}}(n). \tag{44}$$

Similarly, for SDW-MWF, when $\sigma^2_{\text{sd,SDW-MWF}}(n) = K\sigma^2_{X_1}(n)$, according to (38) and (39), the corresponding controlling parameter becomes

$$\lambda_{\text{SDW-MWF}} = \frac{\sqrt{K}\sigma^2_{X_1}(n)}{(1 - \sqrt{K})\sigma^2_{V_{o,\text{MVDR}}}(n)}, \tag{45}$$

and the variance of the residual noise in (40) is given by

$$\sigma^2_{V_{o,\text{SDW-MWF}}}(n) = \left(1 - \sqrt{K}\right)^2 \sigma^2_{V_{o,\text{MVDR}}}(n). \tag{46}$$

Since $\rho < 1$, from (44) and (46), we have

$$\sigma^2_{V_{o,\text{PMKF}}}(n) < \sigma^2_{V_{o,\text{SDW-MWF}}}(n), \tag{47}$$

which means that, for any given level of speech distortion, the PMKF can always yield lower residual noise than the SDW-MWF.

VI. Experiments

We compare the performance of the proposed PMKF with the conventional MVDR beamformer and with the SDW-MWF using a publicly available hearing aid (HA) head-related impulse response (HRIR) database [67].

A. Experimental Setup

With the HRIR database, the eight-channel room impulse responses (RIRs) and the real multichannel noises measured in a cafeteria environment, are used to generate the noisy and reverberant signals received by the microphone array. The eight channels include one in-ear channel and three behind-the-ear channels for each ear, whose geometry is shown in Fig. 1 of [67]. The multichannel noises include the diffuse ambient noise and babble noise, which were recorded during the off-peak and peak times in the same cafeteria environment. The listener is seated at a rectangular table near one corner of the room, and is looking directly at the target speaker seated opposite at a distance of $1.02\,\text{m}$ (position A in Fig. 5 of [68]).

We created $15\,\text{s}$ speech source signals by concatenating randomly selected sentences from the IEEE sentences database [69], and then convolved them with the RIRs to yield the multichannel clean reverberant speech signals. Multichannel noises are added to the clean reverberant signals at certain signal-to-noise ratios (SNRs), which will be elaborated in each experiment, and the factor of scaling multichannel noise is calculated by taking the first channel as reference. All experiments are conducted at a sample rate of 8 kHz. In each trial of the experiment, a new speech signal is randomly generated, and the multichannel clean reverberant speech is added to a randomly selected 15 s interval of the noise signal to obtain the noisy observation.

For all algorithms, we choose the acoustic frame length for the STFT as 16 ms with a 4 ms frame hop, and use a Hamming window. Since the RTF information is utilized by both the MVDR and PMKF, for a fair comparison, based on the discussions in Section V-C2, we implement the SDW-MWF as a concatenation of an MVDR and a Winer-type single-channel post-filter whose filter gain is given by (39). Given the oracle knowledge of the RIRs, the RTF vector is computed using the first 16 ms of the RIRs, and the first channel is taken as reference. The multichannel noise covariance matrix is estimated using the method in [61]. For the PMKF, we set the LP order $P = 2$, and use a modulation frame length of 32 ms with a 16 ms frame hop to estimate the LP coefficients $b_{p,k}$ and the excitation variance $W(n)$ in (4). The pre-processing MWF used in the PMKF is the same as the SDW-MWF with $\lambda = 1$. With the above configurations, each modulation frame consists of 32 ms$/$4 ms $= 8$ acoustic frames. In each frequency bin, based on the pre-processing MWF output, $Z_1(n,k)$, the LP coefficients are recalculated for each new modulation frame given $[Z_1(n,k), Z_1(n-1,k), \ldots, Z_1(n-7,k)]$. These coefficients are kept constant for the 16 ms$/$4 ms $= 4$ acoustic frames within each modulation frame hop.

B. Performance Measure

The perceptual evaluation of speech quality (PESQ) [70] and the frequency-weighted segmental signal-to-noise ratio (FwSegSNR) [71] metrics are used to evaluate the speech enhancement performance of different algorithms. For each metric, we compute the raw value for the reference noisy signal ("$[\cdot]_{\text{raw}}$") and the improvement ("$\Delta[\cdot]$") for the output signal. We average the results over 10 trials.

C. Results With Fixed Controlling Parameter

The controlling parameter λ used for PMKF and SDW-MWF can either be fixed, or be adaptive for each TF bin, to allow the speech enhancement performance to be more flexibly controlled. In this subsection, we will evaluate the algorithms using a fixed controlling parameter, and in the following subsection we evaluate the effect on PMKF and SDW-MWF of adaptively choosing the controlling parameter within each TF bin.

In the fixed controlling parameter case, as we mainly aim to evaluate the effect of the controlling parameter, only two SNRs in the ambient and babble noise conditions are considered, namely -5 dB and 5 dB, and correspond to the highly and moderately noisy environments respectively. The controlling parameter λ changes from -20 dB to 20 dB, with a step size of 4 dB.

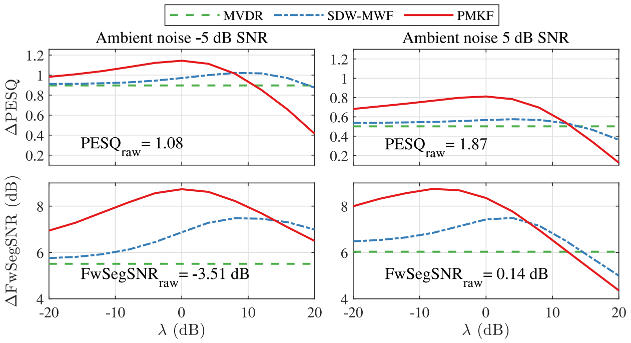

Fig. 3. Comparison results for different values of the controlling parameter λ in ambient noise conditions.

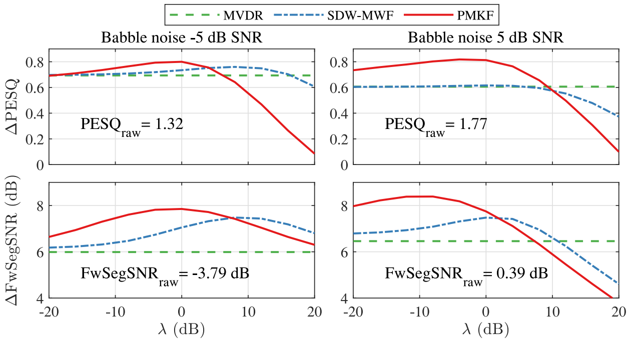

Fig. 4. Comparison results for different values of the controlling parameter λ in babble noise conditions. The PESQ and FwSegSNR of the raw signal are 1.078 and -3.51 dB respectively.

The comparison results for different controlling parameters in ambient and babble noise conditions are shown in Fig. 3 and Fig. 4, respectively. It is shown that when λ is small, both SDW-MWF and PMKF yield similar results to the MVDR beamformer, indicating that in order to control speech distortion, little further noise reduction is performed in SDW-MWF and PMKF. Increasing λ leads to more noise reduction, and this further increases the speech quality and segmental SNR of the SDW-MWF and PMKF outputs.

We can observe that, for a certain range of λ which targets a reasonable trade-off between noise reduction and speech distortion, the proposed PMKF consistently achieves the largest improvement in both PESQ and FwSegSNR for all evaluated noise types and SNRs. When λ is very high, the speech signal is highly distorted, thus both SDW-MWF and PMKF yield worse performances than the MVDR beamformer. However, the results show that, compared with the raw noisy input signal, the SDW-MWF and PMKF still give positive improvements in PESQ and FwSegSNR even for $\lambda = 20$ dB.

It should be noted that both PESQ and FwSegSNR are affected by the noise level. Here we consider one example in the 5 dB ambient noise case to further illustrate the trade-off between speech distortion and noise reduction of SDW-MWF and PMKF. Based on the performance analysis in Section V, the speech distortion variance and residual noise variance metrics of the

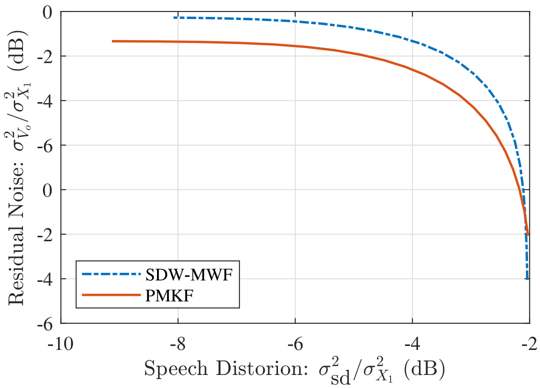

Fig. 5. Illustration of the trade-off behaviour of SDW-MWF and PMKF for one trial in the 5 dB ambient noise condition. The speech distortion variance $\sigma^2_{\rm sd}$ and the noise residual $\sigma^2_{V_o}$ are normalized with respect to variance of the clean signal $\sigma^2_{X_1}$, and the pairwise values for different controlling parameters are shown.

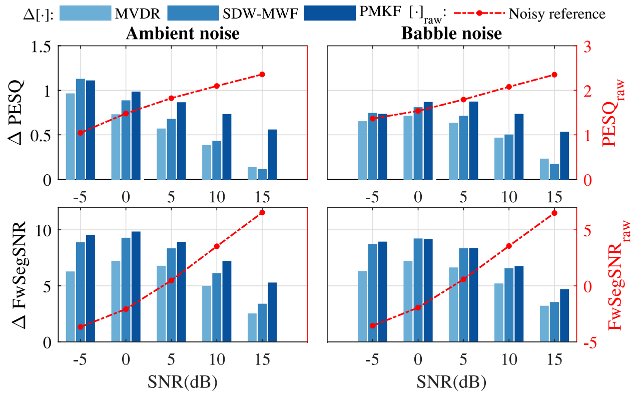

Fig. 6. Comparison results for different SNRs when using the IBM based controlling parameter for SDW-MWF and PMKF. The performance improvements ("Δ[·]") and the raw values ("$[\cdot]_{\rm raw}$") of the reference noisy signal are shown using bars and dashed lines, respectively.

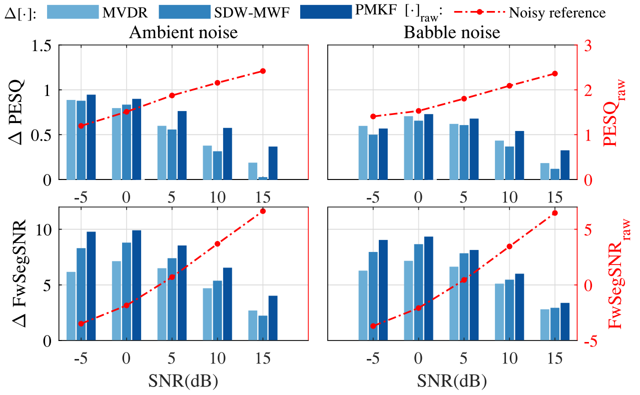

Fig. 7. Comparison results for different SNRs when the controlling parameter for SDW-MWF and PMKF is adaptively chosen according to a binary mask estimated using [73], [74]. The performance improvements ("Δ[·]") and the raw values ("$[\cdot]_{\rm raw}$") of the reference noisy signal are shown using bars and dashed lines, respectively.

PMKF and the SDW-MWF, which are defined in (28), (32), (38) and (40), are calculated in each TF bin. For each controlling parameter, the global variances of speech distortion and residual noise are measured by averaging the metrics over all TF bins in the power domain, and are normalized by the variance of the clean speech. We evaluate the controlling parameter λ from -20 dB to 60 dB with a step size of 4 dB, and the pairwise values of the speech distortion and residual noise measures are plotted in Fig. 5. We can observe for both SDW-MWF and PMKF, decreasing the speech distortion leads to more residual noise, which are consistent with the theoretical analysis in Section V-A, V-B and V-C2. It is shown that to under the same level of speech distortion, the residual noise level of PMKF is smaller than the SDW-MWF.

D. Results With Adaptive Controlling Parameter

In this subsection, we choose the controlling parameter, λ, adaptively for each TF bin. According to the performance analysis in Section V, small controlling parameter values should be used in speech-dominated TF bins to preserve speech, and large values should be used in noise-dominated bins to suppress noise aggressively. Therefore, the choice of controlling parameter depends on the noise level, and thus can be determined based on, for example, the speech presence probability, local SNR, or a mask estimated by neural networks. Alternatively, under the PMKF framework, it is possible to train an end-to-end neural network to find the optimal controlling factor and optimize an overall speech enhancement measure (e.g., MSE) directly.

This section evaluates the use of a TF mask to control the value of λ. The mask is computed either as an ideal binary mask (IBM) [72], or by using the method in [73], [74]. The IBM is computed given oracle knowledge of the speech and noise level in each TF bin. The method in [73], [74] uses the MVDR beamformer output to obtain a ratio mask estimation. The estimated ratio mask is either converted to a binary mask by thresholding at 0.5 or else used directly to compute the controlling parameter.

Given the mask value for each TF bin $\mathcal{M}(n,k)$, the controlling parameter is adaptively chosen as

$$\lambda(n,k) = \mathcal{M}(n,k) * \lambda_{\rm min} + [1 - \mathcal{M}(n,k)] * \lambda_{\rm max}, \quad (48)$$

where the controlling parameter switches or interpolates between the lower bound $\lambda_{\rm min}$ and the upper bound $\lambda_{\rm max}$. In the following experiments, we set $\lambda_{\rm min} = -10\,{\rm dB}$ and $\lambda_{\rm max} = 10\,{\rm dB}$.

We consider the ambient noise and the babble noise cases in the cafeteria environment, and the SNR changes from -5 dB to 15 dB with a step size of 5 dB. In each TF bin, the same controlling parameter is applied on the PMKF and SDW-MWF, and the PESQ and FwSegSNR of the output signal are evaluated.

The comparison results for adaptively choosing the controlling parameter are shown in Fig. 6 to Fig. 8 using respectively the oracle IBM and the ratio mask estimator from [73], [74].

When the IBM based controlling parameters are used, we notice from Fig. 6 that, the proposed PMKF generally achieves the largest improvements in both PESQ and FwSegSNR for all SNRs in ambient noise. In babble noise, although the PMKF

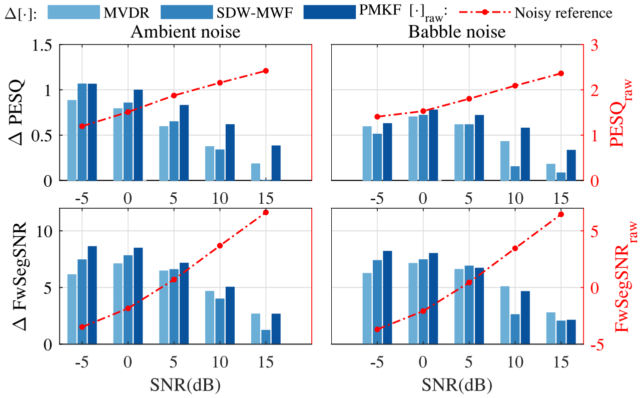

Fig. 8. Comparison results for different SNRs when the controlling parameter for SDW-MWF and PMKF is adaptively chosen according to a ratio mask estimated using [73], [74]. The performance improvements ("$\Delta[\cdot]$") and the raw values ("$[\cdot]_{\mathrm{raw}}$") of the reference noisy signal are shown using bars and dashed lines, respectively.

yields similar improvements to SDW-MWF in FwSegSNR when SNR $\leq$ 10 dB, it still has larger improvements in PESQ than SDW-MWF and MVDR, which indicates that the PMKF outputs have less speech distortion than the SDW-MWF outputs when achieving the similar reduction in noise. It can be also seen that the advantage of PMKF becomes greater as the SNR increases, this is because the preprocessing MWF gives a more accurate estimation of the LP coefficients, and the LP information exploited by the PMKF is more reliable. When SNR $=$ 15 dB, the proposed PMKF gives nearly 0.4 additional improvement in PESQ compared with the SDW-MWF and MVDR outputs. Moreover, since the difference between the noisy and clean signals becomes smaller in high SNR conditions, all methods generally yield decreased improvements over the noisy reference signal as the SNR increases.

The results of using the ratio mask estimator from [73], [74] to determine the controlling parameters without any oracle knowledge are shown in Fig. 7 and Fig. 8. In Fig. 7, the estimated mask is converted to be binary before computing the controlling parameters using (48). Again, we can observe that the proposed PMKF outperforms the SDW-MWF and MVDR for almost all cases in ambient noise. At high SNRs the SDW-MWF fails to improve the PESQ compared with the MVDR, this is possibly because when the estimated mask is not as accurate as the oracle one, the Wiener-like post-filter introduces speech distortion to the relatively clean signal. However, the proposed PMKF can improve both PESQ and FwSegSNR, which indicates that suppressing noise and preserving speech are simultaneously achieved. Similar conclusions can be drawn from Fig. 8 where the ratio mask is directly used in (48) to determine the controlling parameters. According to (48) a smaller controlling parameter is chosen in low SNR TF bins compared with the binary mask case, which makes the noise reduction less aggressive. Therefore, it can be seen that in low SNR conditions, the SDW-MWF and PMKF generally yield a higher PESQ improvement, which indicates that more speech is preserved. It comes at the cost of lower FwSegSNR improvement, meaning that less noise is suppressed.

TABLE I
NORMALIZED EXECUTION TIME OF DIFFERENT ALGORITHMS

Algorithm	MVDR	SDW-MWF	PMKF
Execution time	1	1.52	2.06

E. Computational Complexity Comparison

We finally show the comparison results of computational complexity. We run each algorithm for 50 trials and compare the average execution time of each algorithm. The execution time for modules such as noise covariance matrix estimation and mask estimation is excluded. The result is shown in Table I, which is normalized with respect to that of the MVDR. It can be seen that compared with the MVDR, the computational complexity is increased by 52% by SDW-MWF, and is additionally increased by 36% by the proposed PMKF.

VII. CONCLUSION

A modulation-domain PMKF is proposed by extending the previously proposed MKF and using a parameter to control the trade-off between speech distortion and noise reduction. A new cost function is derived based on the decomposition of the MKF cost function, and the optimal PMKF gain is obtained under the MMSE criterion. We have conducted a performance analysis on the PMKF and the SDW-MWF, and demonstrated that the PMKF can be viewed as integrating the LP information into the SDW-MWF, and the PMKF can always yield lower residual noise than the SDW-MWF with the same amount of speech distortion. Different speech enhancement systems based on various strategies of choosing the controlling parameters are developed in the experiments, and by using a public HRIR database, we demonstrate the effectiveness of the proposed method.

REFERENCES

[1] M. Souden, J. Benesty, and S. Affes, "On optimal beamforming for noise reduction and interference rejection," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2009, pp. 109–112.

[2] 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.

[3] W. Herbordt and W. Kellermann, "Adaptive beamforming for audio signal acquisition," in *Adaptive Signal Processing: Applications to Real-World Problems*, ser. Signals and Communication Technology, J. Benesty and Y. Huang, Eds. Berlin, Germany: Springer-Verlag, 2003, ch. 6, pp. 155–194.

[4] 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.

[5] 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.

[6] 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.

[7] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Proc. 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.

[8] I. Cohen, "Multichannel post-filtering in nonstationary noise environments," *IEEE Trans. Signal Process.*, vol. 52, no. 5, pp. 1149–1160, May 2004.

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

[10] 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.

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

[12] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[13] 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.

[14] 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.

[15] 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.

[16] A. Kuklasinski and J. Jensen, "Multichannel Wiener filters in binaural and bilateral hearing aids-speech intelligibility improvement and robustness to DoA errors," *J. Acoust. Soc. Am.*, vol. 65, no. 1/2, pp. 8–16, 2017.

[17] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," 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.

[18] J. R. Jensen, J. Benesty, and M. G. Christensen, "Variable span filters for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2016, pp. 6505–6509.

[19] J. Benesty, M. G. Christensen, and J. R. Jensen, *Signal Enhancement With Variable Span Linear Filters*. Berlin, Germany: Springer Singapore, 2016.

[20] N. D. Gaubitch, M. R. P. Thomas, and P. A. Naylor, "Dereverberation using LPC-based approaches," in *Speech Dereverberation*, P. A. Naylor and N. D. Gaubitch, Eds., London, U.K.: Springer-Verlag, 2010, pp. 95–128.

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

[22] M. S. Kavalekalam, J. K. Nielsen, J. B. Boldt, and M. G. Christensen, "Model-based speech enhancement for intelligibility improvement in binaural hearing aids," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 99–113, Jan. 2019.

[23] S. So, K. K. Wocicki, J. G. Lyons, A. P. Stark, and K. K. Paliwal, "Kalman fitler with phase spectrum compensation algorithm for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2009, pp. 4405–4408.

[24] Y. Xia and J. Wang, "Low-dimensional recurrent neural network-based kalman filter for speech enhancement," *Neural Netw.*, vol. 67, pp. 131–139, Jul. 2015.

[25] 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.

[26] W.-R. Wu and P.-C. Chen, "Subband Kalman filtering for speech enhancement," *IEEE Trans. Circuits Syst. II: Analog Digit. Signal Process.*, vol. 45, no. 8, pp. 1072–1083, Aug. 1998.

[27] H. Puder, "Speech enhancement with Kalman-filters in subbands," in *Proc. Int. Workshop Acoust. Echo Noise Control.*, Darmstadt, Germany, 2001, pp. 203–207.

[28] H. Puder, "Kalman-filters in subbands for noise reduction with enhanced pitch-adaptive speech model estimation," *Eur. Trans. Telecommun.*, vol. 13, no. 2, pp. 139–148, 2002.

[29] 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.

[30] 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.

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

[41] W. Xue, A. H. Moore, M. Brookes, and P. A. Naylor, "Modulation-domain multichannel Kalman filtering for speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1833–1847, Oct. 2018.

[42] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filters," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006.

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

[44] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction," *Speech Commun.*, vol. 49, no. 7–8, pp. 636–656, Aug. 2007.

[45] D. P. Jarrett, E. A. P. Habets, J. Benesty, and P. A. Naylor, "A tradeoff beamformer for noise reduction in the spherical harmonic domain," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Sep. 2012, pp. 1–4.

[46] K. Ngo, A. Spriet, M. Moonen, J. Wouters, and S. H. Jensen, "Variable speech distortion weighted multichannel wiener filter based on soft output voice activity detection for noise reduction in hearing aids," in *Proc. Int. Workshop Acoust. Signal Enhancement*, vol. 23, 2008.

[47] K. Ngo, A. Spriet, M. Moonen, J. Wouters, and S. Jensen, "Incorporating the conditional speech presence probability in multi-channel Wiener filter based noise reduction in hearing aids," in *Proc. EURASIP J. Adv. Signal Process.*, 2009, pp. 1–11.

[48] K. Ngo, M. Moonen, S. H. Jensen, and J. Wouters, "A flexible speech distortion weighted multi-channel Wiener filter for noise reduction in hearing aids," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2011, pp. 2528–2531.

[49] 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.

[50] W. Xue, A. H. Moore, M. Brookes, and P. A. Naylor, "Modulation-domain parametric multichannel Kalman filtering for speech enhancement," in *Proc. Eur. Signal Process. Conf.*, Sep. 2018, pp. 2509–2513.

[51] O. Schwartz, S. Gannot, and E. A. P. Habets, "Cramér–rao bound analysis of reverberation level estimators for dereverberation and noise reduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 8, pp. 1680–1693, Aug. 2017.

[52] I. Kodrasi and S. Doclo, "Analysis of eigenvalue decomposition-based late reverberation power spectral density estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1106–1118, Jun. 2018.

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

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

[55] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of the direct-path relative transfer function for supervised sound-source localization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2171–2186, Nov. 2016.

[56] 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.

[57] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Am.*, vol. 95, no. 2, pp. 1053–1064, 1994.

[58] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME J. Basic Eng.*, vol. 82, no. 1, pp. 35–45, 1960.

[59] J. D. Markel and A. J. Gray, *Linear Prediction of Speech*. Berlin, Germany: Springer-Verlag, 1976.

[60] 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.

[61] 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.

[62] 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.

[63] M. Brookes, "The matrix reference manual," Imperial College London, 1998–2020. [Online]. Available: http://www.ee.imperial.ac.uk/hp/staff/dmb/matrix/intro.html

[64] M. Brookes, "Inversion identities," Imperial College London, 1998–2020. [Online]. Available: http://www.ee.imperial.ac.uk/hp/staff/dmb/matrix/identity.html#InvLemma

[65] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications.* Berlin, Germany: Springer Science & Business Media, 2001.

[66] L. Ehrenberg, S. Gannot, A. Leshem, and E. Zehavi, "Sensitivity analysis of MVDR and MPDR beamformers," in *Proc. IEEE 26th Conv. Elect. Electron. Eng.*, Israel, Nov. 2010, pp. 416–420.

[67] 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.

[68] R. M. Baumgärtel *et al.*, "Comparing binaural pre-processing strategies I: Instrumental evaluation," *Trends Hear.*, vol. 19, pp. 1–16, 2015.

[69] E. H. Rothauser *et al.*, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, no. 3, pp. 225–246, 1969.

[70] A. W. Rix *et al.*, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2001, pp. 749–752.

[71] 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.

[72] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed., New York, NY, USA: Springer, 2005, pp. 181–197.

[73] A. H. Moore, L. Lightburn, W. Xue, M. Brookes, and P. A. Naylor, "Binaural mask-informed speech enhancement for hearing aids with head tracking," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Sep. 2018, pp. 461–465.

[74] L. Lightburn, "Mask-based enhancement of very noisy speech," Ph.D. dissertation, Imperial College London, 2020.

Wei Xue (Member, IEEE) 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. He has been a Research Scientist with JD AI Research, Beijing, China, since November 2018. From August 2015 to September 2018, he was first a 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 the Universit de Toulon, France, in July 2015, and KU Leuven, Belgium, in September 2016. He was selected into the Beijing Overseas Young Talent Aggregation Project in 2020. His research interest focuses on microphone arrays based speech signal processing, including speech enhancement, sound source localization, and blind system identification.

Alastair H. Moore (Member, IEEE) received the M.Eng. degree in electronic engineering with Music Technology Systems in 2005, and the Ph.D. degree in 2010, both from the University of York, York, U.K. He is a Postdoctoral Researcher with Imperial College London and Spatial Audio Consultant with Square Set Sound. He spent three years as a Hardware Design Engineer with Imagination Technologies PLC designing digital radios and networked audio consumer electronics products. In 2012, he was with Imperial College, where he has contributed to a series of projects in the field of speech and audio processing applied to voice over IP, robot audition, and hearing aids. Particular topics of interest include microphone array signal processing, modeling and characterization of room acoustics, dereverberation, and spatial audio perception. His current research interests include signal processing for moving, and head-worn microphone arrays.

Mike Brookes (Member, IEEE) graduated in mathematics from Cambridge University in 1972. He is a Senior Research Investigator in signal processing with the Department of Electrical and Electronic Engineering , Imperial College London. He was with the Massachusetts Institute of Technology, and briefly, the University of Hawaii before returning to U.K. and joining Imperial College in 1977. Within the area of speech processing, he has concentrated on the modelling 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 (CLEAR) which investigated techniques for processing heavily corrupted speech signals. Between 2015 and 2019, he was Principal Investigator of the E-LOBES project that addressed environment-aware enhancement algorithms for binaural hearing aids.

Patrick A. Naylor (Fellow, IEEE) received the B.Eng. degree in electronic and electrical engineering from the University of Sheffield, U.K., and the Ph.D. degree from Imperial College London, U.K. He is a Professor of Speech and Acoustic Signal Processing with Imperial College London. His research interests include speech, audio and acoustic signal processing. His current research addresses microphone array signal processing, speaker diarization, and multichannel speech enhancement for application to binaural hearing aids and robot audition. He has also worked on speech dereverberation including blind multichannel system identification and equalization, acoustic echo control, non-intrusive speech quality estimation, and speech production modelling with a focus on the analysis of the voice source signal. In addition to his academic research, he enjoys several collaborative links with industry. He is currently a Member of the Board of Governors of the IEEE Signal Processing Society and President of the European Association for Signal Processing (EURASIP). He was formerly Chair of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing. He was an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS and is currently a Senior Area Editor of IEEE TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING.