Modulation-Domain Parametric Multichannel Kalman Filtering for Speech Enhancement

Wei Xue, Alastair H. Moore, Mike Brookes, Patrick A. Naylor
Dept. of Electrical and Electronic Engineering, Imperial College London, UK
Email: {w.xue, alastair.h.moore, mike.brookes, p.naylor}@imperial.ac.uk

***Abstract*—The goal of speech enhancement is to reduce the noise signal while keeping the speech signal undistorted. Recently we developed the multichannel Kalman filtering (MKF) for speech enhancement, in which the temporal evolution of the speech signal and the spatial correlation between multichannel observations are jointly exploited to estimate the clean signal. In this paper, we extend the previous work to derive a parametric MKF (PMKF), which incorporates a controlling factor to achieve the trade-off between the speech distortion and noise reduction. The controlling factor weights between the speech distortion and noise reduction related terms in the cost function of PMKF, and based on the minimum mean squared error (MMSE) criterion, the optimal PMKF gain is derived. We analyse the performance of the proposed PMKF and show the differences with the speech distortion weighted multichannel Wiener filter (SDW-MWF). We conduct experiments in different noisy conditions to evaluate the impact of the controlling factor on the noise reduction performance, and the results demonstrate the effectiveness of the proposed method.**

***Index Terms*—Speech enhancement, Microphone arrays, Kalman filtering, Modulation domain**

I. Introduction

The importance of speech enhancement is widely recognized in recent years. Environmental noise has long been a bottleneck of high performance speech processing systems in robots, hearing aids, mobile devices and smart homes. Multichannel methods are able to capture additional spatial information of the acoustic environment for speech enhancement, and have become a preferred solution compared to the single-channel methods.

The target signal is captured by multiple microphones of a microphone array, and typically includes different spatial information with the noise signal. Examples exist in many geometries including linear and spherical arrays [1]. Multichannel speech enhancement methods generally exploit the spatial diversity between target and noise, and can be categorized as the beamforming based methods [2–5], post-filtering techniques [6–9], and multichannel Wiener filtering (MWF) [10–13]. For these conventional methods, the spatial correlation of the clean signals in different microphones is used to design the optimal filters, however, the temporal correlation of the speech signal is usually neglected.

In our recent work [14], multichannel Kalman filtering (MKF) has been proposed for speech enhancement, which jointly exploits the spatial correlation and temporal evolution of speech. By modelling the speech signal as an auto-regressive (AR) process in the modulation domain, a short-time Fourier transform (STFT)-domain linear prediction (LP) estimation is obtained by first performing LP in the modulation domain, and then inserting the phase from the minimum variance distortion response (MVDR) beamformer output. Based on 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 also shown that the MKF becomes the MWF if the LP information is not incorporated.

Speech enhancement aims to reduce the noise while keeping the speech undistorted. However, it is known that aggressive noise reduction always give rise to speech distortion and, in order to reduce the speech distortion in the output signal, the amount of noise reduction is limited. On one hand, the requirement for noise reduction and speech distortion varies in different applications. On the other hand, it would be beneficial for STFT-domain methods to flexibly control the noise reduction and speech distortion, which, for example, could perform more aggressive noise reduction in speech-absent time-frequency (TF) bins, and limit the speech distortion in speech-present TF bins. In [10–12], the speech distortion weighted MWF (SDW-MWF) has been proposed to achieve a trade-off between speech distortion and noise reduction.

Since it has been shown in [14] that the MKF can be seen as incorporating the LP information into the MWF, it is natural to develop a parametric MKF (PMKF) which has the capability to trade off between the speech distortion and noise reduction. In this paper, the PMKF is proposed which uses a controlling factor to weight between the speech distortion and noise reduction related terms in the cost function of PMKF. Based on the MMSE criterion, the optimal PMKF gain is derived. We conduct experiments in different noisy conditions to evaluate the impact of the controlling factor on the noise reduction performance, and the results demonstrate the effectiveness of the proposed method.

II. Signal Model

We consider a noisy and reverberant environment which includes a single source and an M-element microphone array. The STFT-domain multichannel signal vector in frame n and frequency bin k, $\mathbf{y}(n,k) = [Y_1(n,k)\, Y_2(n,k)\, \ldots\, Y_M(n,k)]^T$, can be expressed as:

$$\begin{aligned}\mathbf{y}(n,k) &= \mathbf{x}(n,k) + \mathbf{v}(n,k) \\ &= \mathbf{d}(k)X_1(n,k) + \mathbf{v}(n,k),\end{aligned} \tag{1}$$

where $\mathbf{x}(n,k)$ and $\mathbf{v}(n,k)$ are defined in a same form as $\mathbf{y}(n,k)$ and denote the clean speech signal vector and additive noise vector, respectively. We assume that the speech and noise signals are uncorrelated. The vector $\mathbf{d}(k) = [D_1(k)\ D_2(k)\ \dots\ D_M(k)]^T$ is the relative transfer function (RTF) between all channels and the reference channel for the target source, and is assumed to be known. Here we take the first channel as reference, therefore, $D_1(k) = 1$.

III. MKF for Speech Enhancement

A. MKF State-Space Model

The MKF proposed in [14] consists of a LP model to describe the temporal evolution of the clean speech signal, and a measurement model to describe the relationship between the clean speech signal and the multichannel noisy observations.

The LP model of MKF is defined on the reference channel:

$$|\mathbf{x}_1(n,k)| = \mathbf{A}(k)|\mathbf{x}_1(n-1,k)| + \mathbf{u}W(n,k), \tag{2}$$

where $\mathbf{x}_1(n,k) = [X_1(n,k)\ X_1(n-1,k)\ \dots\ X_1(n-P+1,k)]^T$ is the signal vector of the first channel, and is also the state vector of MKF. $X_1(n,k)$ is the STFT-domain signal of the first channel. The LP order is P. $\mathbf{A}(k)$ is the speech transition matrix defined in [14], $\mathbf{u} = [1\ 0\ \dots\ 0]^T$ is a $P \times 1$ vector, and $W(n,k)$ is the LP residual with variance δ_W^2.

In practice, $\mathbf{A}(k)$ is unknown and can be estimated via LP analysis in the modulation domain, by using a few acoustic frames of the speech signal of the first channel after noise reduction. Noise reduction is performed using MWF, which is realized by a minimum variance distortion response (MVDR) beamformer with a single-channel Wiener post-filter [16]. We define a $P \times P$ diagonal matrix $\mathbf{\Phi}(n,k)$ whose diagonal elements are the complex exponential phase of $\mathbf{z}_1(n,k)$, where $\mathbf{z}_1(n,k) = [Z_1(n,k)\ Z_1(n-1,k)\ \dots\ Z_1(n-P+1,k)]^H$ is a $P \times 1$ vector of the MWF output.

By incorporating the spatial information, we define the measurement equation using STFT-domain multichannel signals:

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

where $\mathbf{Q}(k) = \mathbf{d}(k)\mathbf{u}^T$ is an $M \times P$ matrix.

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

B. MKF Solution

1) Modulation-domain Linear Prediction: Based on the LP model (2), the amplitude of the state vector in the current frame is estimated in the modulation domain as

$$|\mathbf{x}_1(n|n-1)| = \mathbf{A}|\mathbf{x}_1(n-1|n-1)|, \tag{4}$$

where $\mathbf{x}_1(n|n-1)$ and $\mathbf{x}_1(n-1|n-1)$ are the *a priori* and the *a posteriori* estimates of the current frame and last frame, respectively.

We can further obtain the STFT-domain LP estimation $\mathbf{x}_1(n|n-1)$ by inserting the phase of $\mathbf{z}_1(n)$, then

$$\mathbf{x}_1(n|n-1) = \mathbf{\Phi}(n)|\mathbf{x}_1(n|n-1)|. \tag{5}$$

2) Incorporating Noisy Observation: The state vector is finally updated by combining the estimates from STFT-domain LP and the multichannel noisy observations:

$$\mathbf{x}_1(n|n) = \mathbf{x}_1(n|n-1) + \mathbf{G}(n)[\mathbf{y}(n) - \mathbf{Q}\mathbf{x}_1(n|n-1)], \tag{6}$$

where $\mathbf{G}$ is the MKF gain with dimension $P \times M$.

The error between the $\mathbf{x}_1(n|n)$ and $\mathbf{x}_1(n)$, denoted as $\mathbf{e}(n)$, is computed as

$$\begin{aligned}&\mathbf{e}(n|n)\\ &= \mathbf{x}_1(n|n) - \mathbf{x}_1(n)\\ &= \mathbf{x}_1(n|n-1) - \mathbf{x}_1(n) + \mathbf{G}(n)[\mathbf{y}(n) - \mathbf{Q}\mathbf{x}_1(n|n-1)]\\ &= \mathbf{e}(n|n-1) + \mathbf{G}(n)[\mathbf{Q}\mathbf{e}(n|n-1) + \mathbf{v}(n)]\\ &= [\mathbf{I} - \mathbf{G}(n)\mathbf{Q}]\mathbf{e}(n|n-1) + \mathbf{G}(n)\mathbf{v}(n),\end{aligned} \tag{7}$$

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

$$\mathbf{e}(n|n-1) = \mathbf{x}_1(n|n-1) - \mathbf{x}_1(n). \tag{8}$$

We define an MMSE-based cost function for the MKF as

$$J_{\text{MKF}}[\mathbf{G}(n)] = \text{tr}[\mathbf{R}_{ee}(n|n)], \tag{9}$$

where $\mathbf{R}_{ee}(n|n) = \mathbb{E}\{\mathbf{e}(n|n)\mathbf{e}^H(n|n)\}$. Minimizing $J_{\text{MKF}}[\mathbf{G}(n)]$ leads to the MMSE optimal solution of the MKF gain, $\hat{\mathbf{G}}_{\text{MKF}}(n)$:

$$\begin{aligned}&\hat{\mathbf{G}}_{\text{MKF}}(n)\\ &= \arg\min_{\mathbf{G}(n)} J_{\text{MKF}}[\mathbf{G}(n)]\\ &= \mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H[\mathbf{Q}\mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H + \mathbf{R}_{vv}(n)]^{-1},\end{aligned} \tag{10}$$

where $\mathbf{R}_{ee}(n|n-1) = \mathbb{E}\{\boldsymbol{e}(n|n-1)\boldsymbol{e}^H(n|n-1)\}$ 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 estimation of $\mathbf{R}_{ee}(n|n-1)$ is presented in [14] and $\mathbf{R}_{vv}(n)$ can be estimated by [17–19].

Finally, the clean signal of the first channel $\hat{X}_1(n)$ is estimated as $\mathbf{u}^T\mathbf{x}_1(n|n)$.

IV. Proposed Method

In this section we generalize the MKF to PMKF, which uses a controlling factor to trade off between the speech distortion and noise reduction.

A. Cost function of PMKF

The proposed PMKF adopts the same state-space model as the MKF in Section III-A, and follows (4) and (5) to obtain the STFT-domain LP estimation of the clean signal $\mathbf{x}_1(n|n-1)$.

Since $\mathbf{x}_1(n)$ is the true clean signal, the STFT-domain LP estimation error vector $\mathbf{e}(n|n-1)$ defined in (8), can actually be seen as the speech distortion vector after STFT-domain LP. In (7), with the MKF gain, the speech distortion

vector $\mathbf{e}(n|n-1)$ is converted into one component of $\mathbf{e}(n|n)$, $[\mathbf{I}-\mathbf{G}(n)\mathbf{Q}]\mathbf{e}(n|n-1)$, which can be regarded as the speech distortion after MKF. Because the multichannel noisy observations are incorporated to give the final estimation of the clean signal, $\mathbf{e}(n|n)$ also consists of one noise-related component $\mathbf{G}(n)\mathbf{v}(n)$, which is the residual noise vector in the MKF output.

The $\mathbf{e}(n|n-1)$ is calculated solely from the speech signal, therefore, based on the assumption that the speech and noise are uncorrelated, the MMSE based cost function for MKF in (9), $J_{\text{MKF}}[\mathbf{G}(n)]$, can be rewritten as

$$J_{\text{MKF}}[\mathbf{G}(n)] = \underbrace{\text{tr}\{[\mathbf{I}-\mathbf{G}(n)\mathbf{Q}]\mathbf{R}_{ee}(n|n-1)[\mathbf{I}-\mathbf{G}(n)\mathbf{Q}]^H\}}_{J_s[\mathbf{G}(n)]} + \underbrace{\text{tr}\{\mathbf{G}(n)\mathbf{R}_{vv}(n)\mathbf{G}^H(n)\}}_{J_v[\mathbf{G}(n)]}. \quad (11)$$

Thus the $J_{\text{MKF}}[\mathbf{G}(n)]$ is decomposed into $J_s[\mathbf{G}(n)]$ and $J_v[\mathbf{G}(n)]$, which are related to the speech distortion and the noise residual in the MKF output, respectively.

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

$$J_{\text{PMKF}}[\mathbf{G}_(n)] = J_s[\mathbf{G}(n)] + \lambda J_v[\mathbf{G}(n)], \quad (12)$$

where $\lambda \geq 0$ is the controlling parameter of PMKF. Note that when $\lambda = 1$, the cost function of PMKF is identical to the previously proposed 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. Optimal PMKF Gain

The optimal PMKF gain $\hat{\mathbf{G}}_{\text{PMKF}}(n)$ is obtained by minimizing $J_{\text{PMKF}}[\mathbf{G}(n)]$, based on [20], by setting the derivative of $J_{\text{PMKF}}[\mathbf{G}(n)]$ over $\mathbf{G}(n)$ to zero. We have,

$$\begin{aligned}
&\hat{\mathbf{G}}_{\text{PMKF}}(n) \\
&= \arg\min_{\mathbf{G}(n)} J_{\text{PMKF}}[\mathbf{G}(n)] \\
&= \mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H[\mathbf{Q}\mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H + \lambda\mathbf{R}_{vv}(n)]^{-1} \\
&= \mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H[\mathbf{R}_{Qe}(n) + \lambda\mathbf{R}_{vv}(n)]^{-1}, \quad (13)
\end{aligned}$$

where $\mathbf{R}_{Qe}(n) = \mathbf{Q}\mathbf{R}_{ee}(n|n-1)\mathbf{Q}^H$. From the definition of $\mathbf{Q}$ in (3), $\mathbf{R}_{Qe}(n) = \delta_e^2(n|n-1)\mathbf{d}\mathbf{d}^H$ is of rank-one, where $\delta_e^2(n|n-1) = \mathbf{u}^T\mathbf{R}_{ee}(n|n-1)\mathbf{u}$ is the first diagonal element of $\mathbf{R}_{ee}(n|n-1)$. Therefore a singular matrix inversion problem will occur when $\lambda = 0$ or the noise is absent. The similar issue applies also to (10).

To avoid this problem, in the PMKF, the inverse of matrix $\mathbf{B} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^H$, is computed based on singular value decomposition (SVD) as $\mathbf{V}\bar{\mathbf{\Lambda}}\mathbf{U}^H$, and $\bar{\mathbf{\Lambda}}$ is a diagonal matrix whose i-th diagonal element is $1/\mathbf{\Lambda}_{i,i}$ if $|\mathbf{\Lambda}_{i,i}| > \max_i\{|\mathbf{\Lambda}_{i,i}|\} \times \zeta$, and is 0 otherwise. The $\mathbf{\Lambda}_{i,i}$ is the i-th diagonal element of $\mathbf{\Lambda}$, and in practice, ζ can be chosen according to the uncertainty of the RTF $\mathbf{d}$.

The $\hat{\mathbf{G}}_{\text{PMKF}}(n)$ is substituted into (6), and the clean signal of the first channel is estimated as $\hat{X}_1(n) = \mathbf{u}^T\mathbf{x}_1(n|n)$. The matrix $\mathbf{R}_{ee}(n|n-1)$ is also updated based on $\hat{\mathbf{G}}_{\text{PMKF}}(n)$ as in [14].

C. Analysis

The expression for $\hat{\mathbf{G}}_{\text{PMKF}}(n)$ in (13) is similar to that of the SDW-MWF [11, 21], which is written as

$$\mathbf{h}_{\text{SDW-MWF}} = \mathbf{u}^T\mathbf{R}_{xx}(n)[\mathbf{R}_{xx}(n) + \lambda\mathbf{R}_{vv}(n)]^{-1}, \quad (14)$$

where $\mathbf{R}_{xx}(n) = \mathbb{E}\{\mathbf{x}(n)\mathbf{x}^H(n)\}$ is the speech covariance matrix. However, since the $\mathbf{R}_{ee}(n|n-1)$ in $\hat{\mathbf{G}}_{\text{PMKF}}(n)$ is a function of $\mathbf{e}(n|n-1)$, which depends on the STFT-domain LP estimation $\mathbf{x}_1(n|n-1)$, the speech evolution over time is exploited by the PMKF algorithm but not by SDW-MWF algorithm.

To further elaborate the properties of PMKF, let us consider two extreme cases. Since $\mathbf{u}^T\mathbf{d} = 1$, $\hat{X}_1(n)$ can be expressed as

$$\begin{aligned}
\hat{X}_1(n) &= \mathbf{u}^T\mathbf{d}\mathbf{u}^T\mathbf{x}_1(n|n) \\
&= \mathbf{u}^T\mathbf{Q}\{\hat{\mathbf{G}}_{\text{PMKF}}\mathbf{y}(n) + [\mathbf{I}-\hat{\mathbf{G}}_{\text{PMKF}}\mathbf{Q}]\mathbf{x}_1(n|n)\} \\
&= \mathbf{u}^T\{\mathbf{R}_{Qe}(n)[\mathbf{R}_{Qe}(n) + \lambda\mathbf{R}_{vv}(n)]^{-1}\mathbf{y}(n) + \\
&\quad [\mathbf{Q} - \mathbf{R}_{Qe}(n)(\mathbf{R}_{Qe}(n) + \lambda\mathbf{R}_{vv}(n))^{-1}\times \\
&\quad \mathbf{Q}]\mathbf{x}_1(n|n)\}, \quad (15)
\end{aligned}$$

where "(n)" is omitted for $\hat{\mathbf{G}}_{\text{PMKF}}(n)$ for simplicity.

In the extreme case, when $\lambda = 0$, (15) becomes

$$\begin{aligned}
\hat{X}_1(n) &= \mathbf{u}^T\{\delta_e^2(n|n-1)\mathbf{d}\mathbf{d}^H\frac{\mathbf{d}\mathbf{d}^H}{\|\mathbf{d}\|_2^2\delta_e^2(n|n-1)}\mathbf{y}(n) + \\
&\quad [\mathbf{Q} - \delta_e^2(n|n-1)\mathbf{d}\mathbf{d}^H\frac{\mathbf{d}\mathbf{d}^H}{\|\mathbf{d}\|_2^2\delta_e^2(n|n-1)}\mathbf{Q}]\mathbf{x}_1(n|n)\} \\
&= \mathbf{u}^T\{\frac{1}{\|\mathbf{d}\|_2}\mathbf{d}\mathbf{d}^H\mathbf{y}(n) + [\mathbf{Q} - \frac{1}{\|\mathbf{d}\|_2}\mathbf{d}\mathbf{d}^H\mathbf{Q}]\mathbf{x}_1(n|n)\} \\
&= \frac{1}{\|\mathbf{d}\|_2}\mathbf{d}^H\mathbf{y}(n) + [\mathbf{u}^T - \mathbf{u}^T]\mathbf{x}_1(n|n) \\
&= \frac{1}{\|\mathbf{d}\|_2}\mathbf{d}^H\mathbf{y}(n), \quad (16)
\end{aligned}$$

where $\|\mathbf{d}\|_2 = \mathbf{d}^H\mathbf{d}$. In such case, the PMKF has the form similar to the conventional delay and sum (DS) beamformer [22], therefore, even setting the controlling factor $\lambda = 0$, PMKF achieves signal enhancement. In contrast, the noise will not be suppressed by SDW-MWF, since from (14), $\mathbf{h}_{\text{SDW-MWF}} = \mathbf{u}^T$ if $\lambda = 0$ [11].

When $\lambda = +\infty$, from (15) we have

$$\hat{X}_1(n) = \mathbf{u}^T\mathbf{Q}\mathbf{x}_1(n|n) = \mathbf{u}^T\mathbf{d}\mathbf{u}^T = \mathbf{u}^T\mathbf{x}_1(n|n), \quad (17)$$

which means that the noisy observations are not incorporated and the speech signal is finally estimated as the STFT-domain LP estimation. This property of PMKF is again different with SDW-MWF, which sets the output signal to zero when $\lambda = +\infty$, such that the noise component is totally eliminated.

In summary, when adjusting the controlling parameter of PMKF, the output signal changes between the DS-like beamformer output and the STFT-domain LP estimation. However, in practice, when setting $\lambda = +\infty$, since the STFT-domain LP estimation is computed from the clean signal estimations in previous frames, even though the speech transition matrix $\mathbf{A}$ is non-zero, the LP estimation will approach zero after a period of speech absence and the estimated speech signal gradually vanishes over time.

In addition, we point out that analogous to the Section 3.3 in [14], it can be proved that the PMKF becomes SDW-MWF when the LP information is not adopted. The derivations are omitted in this paper.

V. Experiments

The performance of the proposed PMKF is compared with the conventional MVDR beamformer and the SDW-MWF using a public hearing aid (HA) head-related impluse response (HRIR) database [23]. The MKF and MWF are included as special cases of PMKF and SDW-MWF respectively, by setting $\lambda = 1$ in (13) and (14).

A. Experimental Setup

Six-channel room impulse responses (RIRs) of the HRIR database measured in the cafeteria environment are used to generate the multichannel noisy and reverberant signals with 8 kHz sampling frequency. The six channels include three behind-the-ear (BTE) channels for each ear. The listener is seated at one corner of a rectangle table in the cafeteria, and the target speaker is seated opposite the listener at a distance of 1 m in location 1_A (see Fig. 5 in [24]).

We first obtain a 10 s speech signal by concatenating randomly selected sentences from the IEEE sentences database [25], and then convolve the signal with the listener-specific RIRs to yield the multichannel clean reverberant signal. The multichannel ambient noise and babble noise recorded in the same environment are added to generate the noisy observations. Signal-to-noise ratios (SNRs) of -5 dB and 5 dB are tested.

The STFT frame duration for all algorithms is 16 ms with 4 ms frame hop. The RTF vector is computed using real RIRs truncated to 16 ms with the first channel as reference. We use [17] to estimate the multichannel noise covariance matrix. The LP order of PMKF is $P = 2$, and to estimate the LP coefficients and excitation variance, the modulation-domain frame is 32 ms with 16 ms frame hop. We choose $\zeta = 10^{-6}$. For the SDW-MWF and PMKF, the λ changes from 0.1 to 1000, and the results are shown on a logarithmic scale.

B. Experimental Results

We evaluate the performance using the improvements of short-time objective intelligibility (STOI) [26], perceptual evaluation of speech quality (PESQ) [27], and frequency-weighted segmental SNR (FwSegSNR) [28] metrics over the noisy inputs of the reference channel. Ten trials are conducted and the average results are shown in Fig. 1 and Fig. 2.

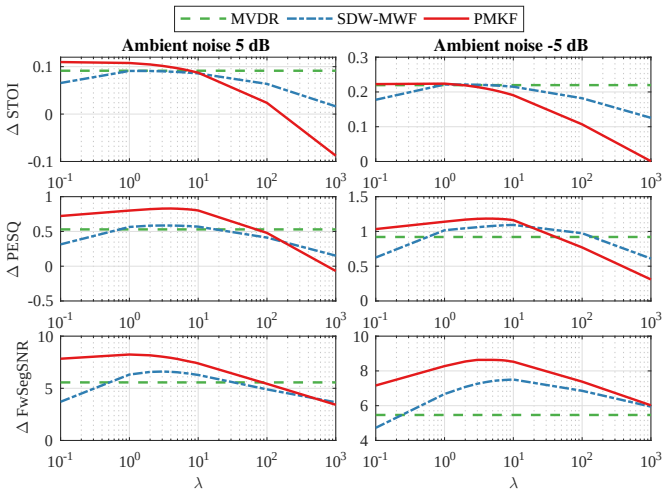

Fig. 1: Comparison results for different values of the controlling parameter λ in ambient noise conditions.

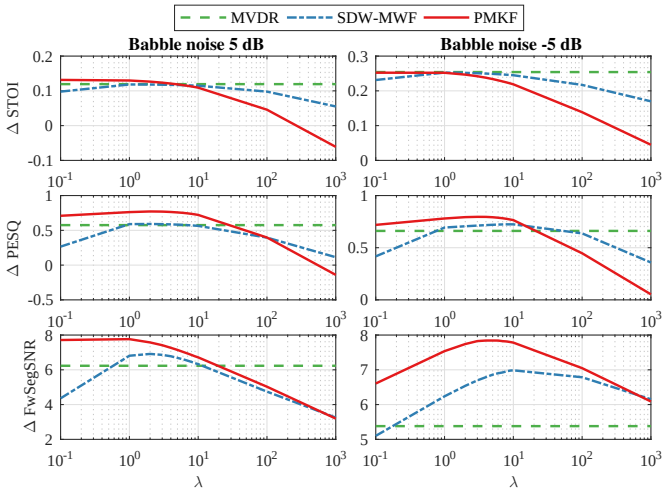

Fig. 2: Comparison results for different values of the controlling parameter λ in babble noise conditions.

It can be seen that when $\lambda < 10$, for all conditions the PMKF yields the largest improvements in PESQ and FwSegSNR, and achieve similar improvements in STOI compared with MVDR. When increasing the λ, especially in high noise scenarios, the FwSegSNR improvement of the PMKF becomes larger, indicating more noise reduction, and the improvement in STOI decreases, as the speech becomes more severely distorted. It can be seen that very large values of λ is not a suitable choice for PMKF, because less speech signal is preserved and therefore making the LP estimation less reliable.

VI. Conclusion

In this paper, we have proposed a PMKF for speech enhancement that uses a controlling factor to trade off between the speech distortion and noise reduction. We derived the optimal PMKF gain based on the MMSE criterion, analysed

the performance of PMKF, and showed the difference between PMKF and SDW-MWF. Simulation results in real-world noisy conditions demonstrate the effectiveness of the proposed method.

ACKNOWLEDGMENT

This work was supported by the Engineering and Physical Sciences Research Council E-LOBES project EP/M026698/1.

REFERENCES

[1] D. P. Jarrett, E. A. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing*. Springer, 2017.

[2] M. Souden, J. Benesty, and S. Affes, "On optimal beamforming for noise reduction and interference rejection," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2009, pp. 109–112.

[3] 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, pp. 158–170, Jan. 2010.

[4] 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.

[5] 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, 2014.

[6] 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.

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

[8] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New York, USA, Apr. 1988, pp. 2578–2581.

[9] I. McCowan and H. Bourlard, "Microphone array post-filter for diffuse noise field," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 2002, pp. 905–908.

[10] 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.

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

[12] S. Doclo, A. Spriet, and M. Moonen, "Efficient frequency-domain implementation of speech distortion weighted multichannel Wiener filtering for noise reduction," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2004, pp. 2007–2010.

[13] 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.

[14] W. Xue, A. H. Moore, M. Brookes, and P. A. Naylor, "Multichannel Kalman filtering for speech enhancement," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[15] F. Gallun and P. Souza, "Exploring the role of the modulation spectrum in phoneme recognition," *Ear and Hearing*, vol. 29, no. 5, p. 800, 2008.

[16] 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, 2016.

[17] 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, 2011.

[18] 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.

[19] 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. Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, Sep. 2012, pp. 1–4.

[20] M. Brookes, "The matrix reference manual," Imperial College London, Website, 1998-2017. [Online]. Available: http://www.ee.imperial.ac.uk/hp/staff/dmb/matrix/intro.html

[21] 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 Communication*, vol. 49, no. 7–8, pp. 636–656, Aug. 2007.

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

[23] 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. on Advances in Signal Processing*, vol. 2009, pp. 298 605:1–10, 2009.

[24] R. M. Baumgärtel, M. Krawczyk-Becker, D. Marquardt, C. Völker, H. Hu, T. Herzke, G. Coleman, K. Adiloglu, S. M. A. Ernst, T. Gerkmann, S. Doclo, B. Kollmeier, V. Hohmann, and M. Dietz, "Comparing binaural pre-processing strategies I: Instrumental evaluation," *Trends in Hearing*, vol. 19, pp. 1–16, 2015.

[25] E. H. Rothauser, W. D. Chapman, N. Guttman, M. H. L. Hecker, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio and Electroacoust.*, vol. 17, no. 3, pp. 225–246, 1969.

[26] 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.

[27] *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, Intl. Telecommunications Union (ITU-T) Recommendation P.862, Feb. 2001.

[28] 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, 2008.