MULTICHANNEL KALMAN FILTERING FOR SPEECH EHNANCEMENT

Wei Xue, Alastair. H. Moore, Mike Brookes, Patrick A. Naylor

Department of Electrical and Electronic Engineering, Imperial College London, UK

ABSTRACT

The use of spatial information in multichannel speech enhancement methods is well established but information associated with the temporal evolution of speech is less commonly exploited. Speech signals can be modelled using an autoregressive process in the time-frequency modulation domain, and Kalman filtering based speech enhancement algorithms have been developed for single-channel processing. In this paper, a multichannel Kalman filter (MKF) for speech enhancement is derived that jointly considers the multichannel spatial information and the temporal correlations of speech. We model the temporal evolution of speech in the modulation domain and, by incorporating the spatial information, an optimal MKF gain is derived in the short-time Fourier transform domain. We also show that the proposed MKF becomes a conventional multichannel Wiener filter if the temporal information is discarded. Experiments using the signals generated from a public head-related impulse response database demonstrate the effectiveness of the proposed method in comparison to other techniques.

Index Terms— Speech enhancement, microphone arrays, Kalman filtering.

1. INTRODUCTION

By using a microphone array, multichannel speech enhancement aims to reduce noise while keeping the target speech signal undistorted. Speech enhancement is of importance in speech processing systems such as speech communication, hearing aids and automatic speech recognition.

Conventional multichannel speech enhancement algorithms include beamforming [1–3], multichannel Wiener filtering (MWF) [4, 5], and post-filtering techniques [6, 7]. Because the target speech signal is redundantly expressed in multiple microphones up to a relative transfer function (RTF), which describes the spatial correlation of multiple clean signals, optimal filters can be derived that combine the multichannel noisy observations to estimate the target clean speech. Although these methods commonly use the spatial correlations present in multichannel signals for speech enhancement, they normally ignore the temporal correlation that speech signals also exhibit.

The temporal evolution of speech can be modelled as an autoregressive (AR) process in the modulation domain. By using linear prediction (LP) to model the temporal correlation of speech, single-channel Kalman filtering (KF) based methods [8–15], especially the modulation-domain ones [10–15], have shown performance superior to conventional methods. In the multichannel case, due to the spatial and temporal correlation, the clean speech signal of one channel is related both to the signals of other channels and to the signals of previous time-frames. This motivates the development of multichannel Kalman filtering (MKF) for speech enhancement.

In this paper, we aim to make the MKF combine the spatial information from multichannel observations with the temporal correlation of speech. However, a difficulty is that the temporal evolution of speech is reflected in the modulation domain, while the spatial information is contained in the complex short-time Fourier transform (STFT) domain. A novel approach which exploits both domains is proposed, and an STFT-domain-optimal MKF gain is derived to determine the weight between the STFT-domain LP estimate and the estimate from noisy observations. We demonstrate that the proposed MKF becomes the conventional MWF if the LP estimation, which captures the temporal correlations, is discarded. This shows that, the proposed MKF additionally takes advantage of temporal correlations of speech compared with standard approaches employing the MWF. We show in this paper the development of our method, relevant analyses and a performance evaluation using the signals generated by a public head-related impulse response (HRIR) database [16].

2. SIGNAL MODEL

Given an M-element microphone array in a reverberant and noisy environment, the $M \times 1$ multichannel signal vector in the STFT domain, $\mathbf{y}(n,k) = [Y_1(n,k)\, Y_2(n,k)\, \ldots\, Y_M(n,k)]^T$, for the n-th frame and k-th frequency bin is written as

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

where $\mathbf{x}(n,k)$ and $\mathbf{v}(n,k)$ are the reverberant clean speech and additive noise vectors, respectively, and have the same

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

form as $\mathbf{y}(n,k)$. $X_1(n,k)$ represents the clean signal of the first channel. The RTF between all channels and the reference channel, $\mathbf{h}(k) = [H_1(n,k)\ H_2(n,k)\ \dots\ H_M(n,k)]^T$, is assumed to be known a priori or to have been already estimated (e.g. as in [17]). Without loss of generality, we take the first channel as reference, and therefore $H_1(n,k) = 1$. Our goal is to estimate the clean reverberant signal $X_1(n,k)$ based on the multichannel noisy observations $\mathbf{y}(n,k)$. We assume that the speech and noise signals are uncorrelated.

3. PROPOSED METHOD

3.1. MKF System Model

First we model the temporal evolution of speech. In the first step, we only consider the signal of the reference channel and ignore the spatial information. Since the temporal evolution is mainly reflected in the magnitude spectrum, we formulate the LP equation of MKF in the modulation domain as

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

where $|\cdot|$ takes the magnitude of each element of the vector. $\mathbf{x}_1(n,k) = [X_1(n,k)\ X_1(n-1,k)\ \dots\ X_1(n-P+1,k)]^T$ is the temporal signal vector of the first channel, and forms the state vector of the MKF. $\mathbf{A}(k)$ is the speech transition matrix defined as $\begin{bmatrix} -\mathbf{a}^T(k) \\ \mathbf{I}_{(P-1)\times(P-1)} \quad \mathbf{0}_{(P-1)\times 1} \end{bmatrix}$, where $\mathbf{a}(k) = [a_{1,k}\ \dots\ a_{P,k}]^T$ is the LP coefficient vector of order P. The vector $\mathbf{d} = [1\ 0\ \dots\ 0]^T$ is a $P \times 1$ vector, $W(n,k)$ is a random Gaussian excitation signal with variance δ_W^2.

In practice, $\mathbf{a}(k)$ and δ_W^2 are unknown and must be estimated via LP analysis of a pre-cleaned speech signal, $\mathbf{z}_1(n,k)$, from modulation frames as in, for example, [10]. We take $\mathbf{z}_1(n,k)$ as the output of a MWF, which is realized as a minimum variance distortion response (MVDR) beamformer followed by a single-channel Wiener post-filter [18]. The phase of $\mathbf{z}_1(n,k)$ is the same with the MVDR output.

We now incorporate the spatial information to define the measurement equation. To preserve the spatial information mainly carried by the phase spectrum, from (1), the measurement equation is defined in the complex STFT domain:

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

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

3.2. Derivation of MKF

Since the LP equation and measurement equation are defined in the modulation domain and STFT domain respectively, the state vector cannot be estimated in the conventional KF framework. In this subsection, an MKF is derived to estimate the state vector that represents the target clean signal.

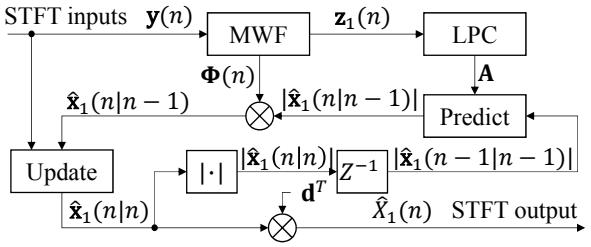

Fig. 1: MKF framework. The frequency index k is omitted for brevity.

The framework of the proposed MKF is illustrated in Fig. 1. Similar to the conventional KF, the proposed MKF consists of a LP step and an update step. In the LP step, an *a priori* modulation-domain estimate $|\hat{\mathbf{x}}_1(n|n-1,k)|$ is calculated and this is then transformed into the STFT domain $\hat{\mathbf{x}}_1(n|n-1,k)$ by incorporating the phase of the MWF output. In the update step, by incorporating the noisy observation, the optimal MKF gain is derived, and used to compute the *a posteriori* state vector estimation $\hat{\mathbf{x}}_1(n|n,k)$. Details of the MKF framework will be described in the following.

3.2.1. STFT-domain State Vector Prediction

Given the MMSE estimate of the state vector of the previous frame, $\hat{\mathbf{x}}_1(n-1|n-1,k)$, the amplitude of the state vector in the current frame can be predicted using (2) as

$$|\hat{\mathbf{x}}_1(n|n-1,k)| = \mathbf{A}(k)|\hat{\mathbf{x}}_1(n-1|n-1,k)|. \quad (4)$$

From $|\hat{\mathbf{x}}_1(n|n-1,k)|$, we can further obtain the STFT-domain LP estimate $\hat{\mathbf{x}}_1(n|n-1,k)$ by inserting the phase of $\mathbf{z}_1(n,k)$, such that

$$\hat{\mathbf{x}}_1(n|n-1,k) = \mathbf{\Phi}(n,k)|\hat{\mathbf{x}}_1(n|n-1,k)|, \quad (5)$$

where $\mathbf{\Phi}(n,k)$ is a $P \times P$ diagonal matrix whose diagonal elements are the complex exponential phase of $\mathbf{z}_1(n,k)$.

We define the STFT-domain LP estimation error between $\mathbf{x}_1(n,k)$ and $\hat{\mathbf{x}}_1(n|n-1,k)$, which will be used later, as

$$\boldsymbol{e}(n|n-1,k) = \hat{\mathbf{x}}_1(n|n-1,k) - \mathbf{x}_1(n,k). \quad (6)$$

3.2.2. STFT-domain State Vector Update

For each new frame, MKF updates the state vector by combining the estimates from LP, $\hat{\mathbf{x}}_1(n|n-1,k)$, and the new observation, $\mathbf{y}(n,k)$, so that

$$\begin{aligned} \hat{\mathbf{x}}_1(n|n,k) = \hat{\mathbf{x}}_1(n|n-1,k) \\ + \mathbf{G}(n,k)[\mathbf{y}(n,k) - \mathbf{Q}(k)\hat{\mathbf{x}}_1(n|n-1,k)], \end{aligned} \quad (7)$$

where $\mathbf{G}(n,k)$ is the MKF gain.

Then the error between $\mathbf{x}_1(n,k)$ and the new estimate $\hat{\mathbf{x}}_1(n|n,k)$ is computed as

$$\begin{aligned}
&\boldsymbol{e}(n|n,k)\\
&= \hat{\mathbf{x}}_1(n|n,k) - \mathbf{x}_1(n,k)\\
&= \hat{\mathbf{x}}_1(n|n-1,k) - \mathbf{x}_1(n,k)\\
&\quad + \mathbf{G}(n,k)[\mathbf{y}(n,k) - \mathbf{Q}(k)\hat{\mathbf{x}}_1(n|n-1,k)]\\
&= \boldsymbol{e}(n|n-1,k) + \mathbf{G}(n,k)[\mathbf{v}(n,k) - \mathbf{Q}(k)\boldsymbol{e}(n|n-1,k)]\\
&= [\mathbf{I} - \mathbf{G}(n,k)\mathbf{Q}(k)]\boldsymbol{e}(n|n-1,k) + \mathbf{G}(n,k)\mathbf{v}(n,k).
\end{aligned} \tag{8}$$

To obtain the optimal MKF gain, we define a cost function under the MMSE criterion as

$$J(\mathbf{G}(n,k)) = \mathbb{E}\{\mathrm{tr}[\boldsymbol{e}(n|n,k)\boldsymbol{e}^H(n|n,k)]\}. \tag{9}$$

Setting the derivative of $J(\mathbf{G}(n,k))$ over $\mathbf{G}(n,k)$ [19] to zero,

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

where $\mathbf{R}_{ee}(n|n-1,k) = \mathbb{E}\{\boldsymbol{e}(n|n-1,k)\boldsymbol{e}^H(n|n-1,k)\}$ is the covariance matrix of the STFT-domain estimation error, $\mathbf{R}_{vv}(n,k) = \mathbb{E}\{\mathbf{v}(n,k)\mathbf{v}^H(n,k)\}$ is the multichannel noise covariance matrix which can be estimated using [20–22].

The matrix $\mathbf{R}_{ee}(n|n-1,k)$ is unknown and will be estimated in the next subsection. We note that since $\mathbf{Q}(k)\mathbf{R}_{ee}(n|n-1,k)\mathbf{Q}^H(k)$ is of rank-one according to the definition of $\mathbf{Q}(k)$ in (3), the matrix inverse in (10) must be replaced by the pseudo-inverse if $\mathbf{R}_{vv}(n,k) = \mathbf{0}$.

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

Assuming from (5) the phase of $\hat{\mathbf{x}}_1(n|n-1,k)$, which is the same as that of the MVDR output, is a good approximation of the phase of $\mathbf{x}_1(n,k)$. Then $\mathbf{x}_1(n,k)$ can be rewritten as $\mathbf{x}_1(n,k) = \boldsymbol{\Phi}(n,k)|\mathbf{x}_1(n,k)|$. From (6), we can obtain

$$\begin{aligned}
\boldsymbol{e}(n|n-1,k) &= \boldsymbol{\Phi}(n,k)[|\mathbf{x}_1(n,k)| - |\hat{\mathbf{x}}_1(n|n-1,k)|]\\
&= \boldsymbol{\Phi}(n,k)\boldsymbol{\epsilon}(n|n-1,k),
\end{aligned} \tag{11}$$

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

Define $\mathbf{R}_{\epsilon\epsilon}(n|n-1,k) = \mathbb{E}\{\boldsymbol{\epsilon}(n|n-1,k)\boldsymbol{\epsilon}^H(n|n-1,k)\}$ as the covariance matrix of the modulation-domain estimation error. It can be updated based on the conventional single-channel modulation-domain KF [10], as

$$\begin{aligned}
\mathbf{R}_{\epsilon\epsilon}(n|n-1,k) &= \mathbf{A}(k)\mathbf{R}_{\epsilon\epsilon}(n-1|n-1,k)\mathbf{A}^H(k)\\
&\quad + \delta_W^2\mathbf{d}\mathbf{d}^H.
\end{aligned} \tag{12}$$

From (11), by considering the phase information, $\mathbf{R}_{ee}(n|n-1,k)$ in (10) can be computed as

$$\mathbf{R}_{ee}(n|n-1,k) = \boldsymbol{\Phi}(n,k)\mathbf{R}_{\epsilon\epsilon}(n|n-1,k)\boldsymbol{\Phi}^H(n,k). \tag{13}$$

Algorithm 1: MKF

$\hat{\mathbf{x}}_1(n|n,k) = [Y_1(n,k)\ \ldots\ Y_1(n-P+1,k)]^H, n \le P.$
for $n = P+1$ to N **do**
  a) Determine $|\hat{\mathbf{x}}_1(n|n-1,k)|$ and $\mathbf{R}_{\epsilon\epsilon}(n|n-1,k)$ from (4) and (12);
  b) Determine $\hat{\mathbf{x}}_1(n|n-1,k)$ and $\mathbf{R}_{ee}(n|n-1,k)$ from (5) and (13);
  c) Compute the MKF gain $\mathbf{G}(n,k)$ from (10);
  d) Update the state vector $\hat{\mathbf{x}}_1(n|n,k)$ using (7);
  e) Update $\mathbf{R}_{ee}(n|n,k)$ and $\mathbf{R}_{\epsilon\epsilon}(n|n,k)$ from (14) and (15);
end
N is the number of frames.

The updating steps in (12) and (13) are from the prediction step. After incorporating the noisy observation, substituting (10) into (8), similarly to the conventional KF, we finally have

$$\mathbf{R}_{ee}(n|n,k) = [\mathbf{I} - \hat{\mathbf{G}}(n,k)\mathbf{Q}(k)]\mathbf{R}_{ee}(n|n-1,k). \tag{14}$$

Note that $\boldsymbol{\Phi}^{-1}(n,k) = \boldsymbol{\Phi}^H(n,k)$, so

$$\mathbf{R}_{\epsilon\epsilon}(n|n,k) = \boldsymbol{\Phi}^H(n,k)\mathbf{R}_{ee}(n|n,k)\boldsymbol{\Phi}(n,k). \tag{15}$$

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

3.3. Relationship with MWF

If the estimates from LP are not used, by setting $\hat{\mathbf{x}}_1(n|n-1,k) = 0$, $\boldsymbol{e}(n,k)$ in (6) becomes $-\mathbf{x}_1(n,k)$, and $\mathbf{Q}(k)\boldsymbol{e}(n,k)$ becomes $-\mathbf{x}(n,k)$. Then $\mathbf{Q}(k)\mathbf{R}_{ee}(n|n-1,k)\mathbf{Q}^H(k)$ in (10) turns into the speech covariance matrix $\mathbf{R}_{xx}(n,k) = \mathbb{E}\{\mathbf{x}(n,k)\mathbf{x}^H(n,k)\}$. Note that $\mathbf{d}^T\mathbf{h}(k) = 1$, then

$$\begin{aligned}
\mathbf{d}^T\hat{\mathbf{x}}_1(n|n,k) &= \mathbf{d}^T\mathbf{h}(k)\mathbf{d}^T\hat{\mathbf{x}}_1(n,k)\\
&= \mathbf{d}^T\mathbf{Q}(k)\hat{\mathbf{G}}(n,k)\mathbf{y}(n,k)\\
&= \mathbf{d}^T\mathbf{R}_{xx}(n,k)\times\\
&\quad [\mathbf{R}_{xx}(n,k) + \mathbf{R}_{vv}(n,k)]^{-1}\mathbf{y}(n,k).
\end{aligned} \tag{16}$$

Thus the MKF becomes the MWF in [4]. Therefore, the proposed MKF can be seen as integrating the temporal evolution of speech into the conventional MWF.

4. EXPERIMENTAL EVALUATION

4.1. Experimental Setup

Algorithms are tested using the cafeteria environment in the HRIR database [16], which is represented by measured RIRs for a pair of behind-the-ear hearing aids, each with three microphones. The target speaker is seated opposite the listener in the

look direction (RIR 1_A in [16]) while the optional interferer is to the left of the listener (RIR 1_C in [16]).

A 10 s speech signal with 8 kHz sampling rate is first generated from randomly selected sentences of the IEEE sentences database [23], and then convolved with RIR 1_A to yield the clean reverberant speech. Algorithms are tested both with and without an interfering source. For the case of no interferer, noisy reverberant signals are generated by adding multichannel ambient noise or babble noise recorded in the same room, at signal-to-noise ratios (SNRs) ranging from -5 dB to 15 dB. For the with-interferer case, the SNR of the babble noise is fixed at 10 dB, and the signal-to-interference ratio (SIR) ranges from -5 dB to 15 dB. The interferer signal was either speech shaped noise (SSN) or white Gaussian noise (WGN).

The proposed method is compared with the widely used MVDR and MWF algorithms. Since the proposed MKF and MVDR exploit the RTF, the MWF is, for fair comparison, implemented by concatenating an MVDR beamformer and single-channel Wiener filter, which can generally achieve better performance than the implementation [4] without using the RTF. 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, and the first channel is taken as the reference. The multichannel noise covariance matrix is estimated using [20]. In the proposed MKF, the LP order $P = 2$, and the LP coefficients and excitation variance are estimated using 8 acoustic frames with a hop of 4 acoustic frames.

4.2. Experimental Results

The performance is assessed with the short-time objective intelligibility (STOI) [24], perceptual evaluation of speech quality (PESQ) [25], and frequency-weighted segmental SNR (FwSegSNR) [26] metrics. For each metric, by taking the raw noisy input of the first channel as reference, we present both the improvement ("$\Delta[\cdot]$") and the raw value ("$[\cdot]_{\text{raw}}$") of the reference signal. The results are averaged over 10 trials.

The results without any interfering source are shown in Fig. 2a. The results show that the algorithms have similar improvement in STOI, and the proposed method achieves substantially larger improvements than other methods in PESQ and FwSegSNR. Although the MWF outperforms the MVDR in PESQ and FwSegSNR, when SNR $\leq$ 10 dB, by exploiting the temporal evaluation of speech, the MKF yields at least 0.095 improvement in PESQ and 0.83 dB improvement in FwSegSNR relative to the MWF. It is also seen that the improvements in all metrics decrease when the SNR increases. This is because in such conditions the difference between the clean and noisy speech becomes smaller, then there is less room for improvement.

Similar results are shown in Fig. 2b for the with-interference scenarios. Again, the proposed MKF always achieves the best performance in ΔPESQ and ΔFwSegSNR, and has the greatest ΔSTOI in almost all cases. It can also be observed

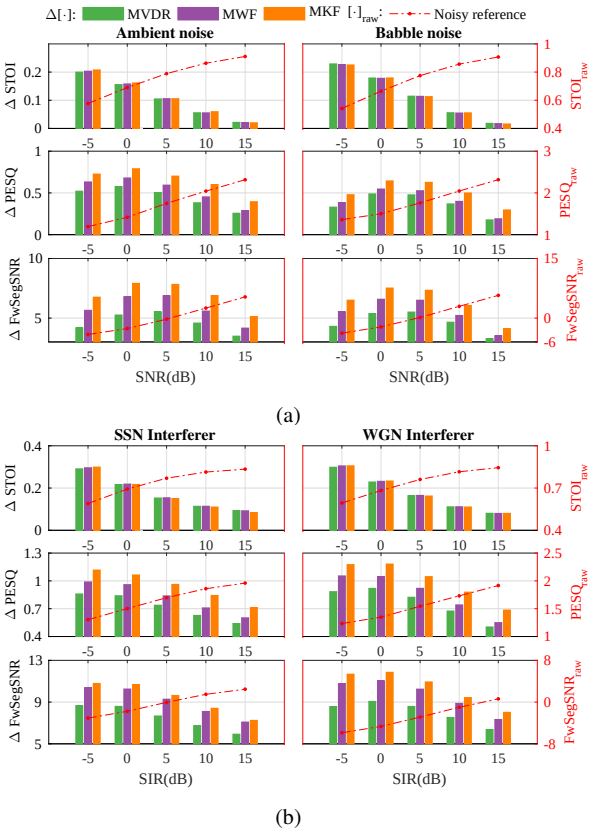

Fig. 2: Comparison results for (a) different SNRs and noise types without an interfering source, (b) different SIRs and interferering source types with 10 dB SNR babble noise always present. The mean metric of the raw signals is shown as a dash-dot line using the rightmost axis; the improvements obtained from the algorithms are shown in bars using the leftmost axis.

that the improvements in the WGN case is generally greater than those in the SSN case, since the spectrum of SSN is more similar to that of the speech signal, making the speech signal more likely to be affected by noise.

5. CONCLUSION

An MKF based speech enhancement method has been presented. The key feature of the method is jointly exploiting both spatial and temporal information simultaneously in the design of the multichannel speech enhancement procedure in the context of MKF. By performing LP in the modulation domain and incorporating the spatial information in the STFT domain, an optimal MKF gain is derived to compromise between the LP estimation and observation adaptively. It has been shown that the MKF becomes the MWF when the LP estimate is not used. Experiments show that the MKF outperforms conventional methods in a range of noisy reverberant conditions.

6. REFERENCES

[1] B. D. van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoustics, Speech and Signal Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[2] 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.

[3] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Relaxed binaural LCMV beamforming," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 137–152, jan 2017.

[4] 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.

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

[6] 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.

[7] I. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 709–716, Nov. 2003.

[8] T. Esch and P. Vary, "Speech enhancement using a modified Kalman filter based on complex linear prediction and supergaussian priors," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 4877–4880.

[9] H. Puder, "Kalman-filters in subbands for noise reduction with enhanced pitch-adaptive speech model estimation," *European Transactions on Telecommunications*, vol. 13, no. 2, pp. 139–148, 2002.

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

[11] Y. Wang and M. Brookes, "Speech enhancement using a robust Kalman filter post-processor in the modulation domain," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, May 2013, pp. 7457–7461.

[12] ——, "Speech enhancement using a modulation domain Kalman filter post-processor with a Gaussian mixture noise model," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7024–7028.

[13] ——, "Speech enhancement using an MMSE spectral amplitude estimator based on a modulation domain Kalman filter with a Gamma prior," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5225–5229.

[14] 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 on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2017, pp. 111–115.

[15] ——, "Speech enhancement using modulation-domain Kalman filtering with active speech level normalized log-spectrum global priors," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2017.

[16] 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.

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

[18] 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.

[19] 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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] *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.

[26] 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.