JOINT BEAMFORMING AND REVERBERATION CANCELLATION USING A CONSTRAINED KALMAN FILTER WITH MULTICHANNEL LINEAR PREDICTION

Sahar Hashemgeloogerdi[1], *Sebastian Braun*[2]

[1] Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY, USA
[2] Microsoft Research, Redmond, WA, USA
shashemg@ur.rochester.edu, sebastian.braun@microsoft.com

ABSTRACT

The performance of speech processing systems degrades significantly in far-field scenarios where the distance between the user and microphones increases, leading to low signal-to-noise and signal-to-reverberation ratios. To address this challenge, combining the denoising and dereverberation techniques in both parallel and cascade configurations has been widely studied. However, a parallel or cascade combination may not be efficient while imposing a large computational complexity. We propose a constrained Kalman filter based multichannel linear prediction method to jointly perform denoising and dereverberation efficiently using an online processing algorithm. In contrast to previously proposed methods which utilize steering vectors based on the relative early transfer function, our algorithm is implemented using a direct relative transfer function based steering vector, which aims at extracting the direct sound as opposed to preserving the early reflections. We show that the proposed algorithm outperforms existing online implementations of integrated beamformer and linear prediction methods on the REVERB challenge speech enhancement task while being computationally less complex.

Index Terms— Dereverberation, denoising, constrained Kalman filter, multichannel linear prediction.

1. INTRODUCTION

The performance of speech processing applications, e.g., hands-free teleconferencing systems and automatic speech recognition and identification systems significantly decreases in far-field scenarios where the microphone is located at far distances from the sound source. This causes a significant distortion in speech quality and intelligibility [1, 2]. Microphone arrays, which allow spatial filtering of the arriving signals, have been employed to suppress interfering sound and enhance the desired signal. Statistically optimal beamformers such as minimum variance distortionless response (MVDR) [3], linearly constrained minimum variance (LCMV) [4, 5], MVDR with a constrained Kalman filter [6, 7] and generalized sidelobe canceler (GSC) [8] have been widely utilized for speech extraction in noisy environments. The performance of linear spatial filters is limited in suppressing reverberation, which is typically time-varying and spatially diffuse. Dereverberation methods based on multichannel linear prediction (MCLP) have been shown to be highly effective [9, 10]. In the MCLP model, the reverberation is represented by an autoregressive model, and can therefore be predicted from previous frames of the microphone signals [9, 10].

When considering also additive noise in MCLP model, finding the correct dereverberation filter becomes a more complex problem. Joint denoising and dereverberation methods based on the MCLP model have been proposed in [2, 11]. Furthermore, combinations of beamforming with MCLP based dereverberation methods to suppress both reverberation and noise have been proposed by cascaded systems in [11–14]. To overcome the possibly inefficient cascade [12], unified approaches have recently been proposed, e. g., integrated sidelobe cancellation and linear prediction (ISCLP) [15], weighted power minimization distortionless response beamformer (WPD) [16], and recursive WPD and recursive least-squares WPD (RLS-WPD) methods [17]. The online processing version WPD-RLS [17] is similar to the ISCLP structure by using a GSC beamformer. While the ISCLP and WPD methods reduce the computational complexity compared to the cascade configuration, they are shown to perform on par or better than the cascade [15, 17]. Furthermore, the ISCLP and WPD methods both utilize steering vectors based on the relative early transfer functions (RETFs), which are estimated either by computationally rather complex methods using the generalized eigenvalue decomposition method [5, 18] or a neural network-based mask [17, 19, 20].

In this paper, we propose an algorithm for joint beamforming and dereverberation using a constrained Kalman filter with the MCLP signal model. This algorithm may overcome the problem of signal distortion along the look direction which typically occurs in GSC-based algorithms. While we found that the GSC based ISCLP suffers from heavy speech cancellation when using direct relative transfer functions (DRTFs), which seemed to motivate the choice of using RETFs, we show that our proposed algorithm is robust enough to use either DRTFs or RETF. The proposed overall system can be implemented with less complexity than previously proposed methods, as any simple geometry-based direction-of-arrival (DOA) estimator can be used to obtain the steering vector, and additionally, the constrained Kalman filter yields slightly less complexity than the GSC [15] due to the absence of a blocking matrix. We evaluate the proposed method in noisy and reverberant environments using the REVERB challenge dataset [21, 22]. Experimental results show that our proposed method outperforms comparable state-of-the-art methods.

The rest of the paper is organized as follows. In Section 2, we define the model of the signal. In Section 3, we describe the ISCLP. We present a derivation of the proposed method in Section 4. We then evaluate the performance of our proposed algorithm for speech enhancement systems in Section 5. Conclusions are presented in Section 6.

Sahar Hashemgeloogerdi performed this work as an intern while at Microsoft Research.

2. SIGNAL MODEL

We assume that M microphones capture the sound in a reverberant and noisy environment. The m-th microphone signal in the short-time Fourier transform (STFT) domain is denoted by $y_m(k,n)$, where k and n are the frequency and time indices, respectively. We describe the vector of microphone signals as $\mathbf{y}(k,n) = [y_1(k,n), \cdots, y_M(k,n)]^T$, which can be formulated as

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

where $x(k,n)$ is the desired signal at the reference microphone, $\mathbf{g}(k,n)$ is a relative transfer function, and $\mathbf{r}(k,n)$ and $\mathbf{v}(k,n)$ denote the late reverberation and additive noise, respectively. The frequency index k is omitted in the rest of the paper for better readability. The vector $\mathbf{g}(n)$ can either represent the DRTFs or the RETFs, which changes the desired signal $x(n)$ accordingly to represent either the direct sound at the reference microphone, or to contain also early reflections. We assume that the desired speech signal $x(n)$ has a zero-mean complex Gaussian distribution

$$x(n) \sim \mathcal{N}(0, \Phi_x(n)), \tag{2}$$

where $\Phi_x(n)$ is the power spectral density (PSD) of $x(n)$. We describe the late reverberation using the MCLP model [23] as a delayed prediction by D from the past L frames by

$$\mathbf{r}(n) = \sum_{l=D}^{L} \mathbf{W}_{\mathrm{r},l}(n)\mathbf{y}(n-l), \tag{3}$$

where $\mathbf{W}_{\mathrm{r},l}(n)$ denotes the MCLP coefficients, and $L > D > 1$.

3. REVIEW OF INTEGRATED SIDELOBE CANCELLATION AND LINEAR PREDICTION

In this section, we review the ISCLP method which integrates a GSC beamformer structure, where the sidelobe canceler aims at cancelling the noise at the fixed beamformer output, and a MCLP branch aiming at estimating the reverberation at the output of the beamformer [15]. The sidelobe canceler branch is given by

$$\mathbf{u}_{\mathrm{SC}}(n) = \mathbf{B}^H(n)\mathbf{y}(n), \tag{4}$$

where $\mathbf{B}(n)$ is a blocking matrix such that $\mathbf{B}^H(n)\mathbf{g}(n) = \mathbf{0}$. The output signal of the ISCLP framework is given by

$$\begin{aligned} e_{\mathrm{ISCLP}}(n) = \mathbf{w}_{\mathrm{g}}^H(n)\mathbf{y}(n) - \mathbf{w}_{\mathrm{SC}}^H(n)\mathbf{u}_{\mathrm{SC}}(n) \\ - \sum_{l=D}^{L} \mathbf{w}_{\mathrm{LP}}^H(l,n)\mathbf{y}(n-l), \end{aligned} \tag{5}$$

where $\mathbf{w}_{\mathrm{g}} = \mathbf{g}/\|\mathbf{g}\|_2^2$ are fixed beamformer coefficients, $\mathbf{w}_{\mathrm{SC}}(n)$ is the SC filter aiming at cancelling noise, and $\mathbf{w}_{\mathrm{LP}}(l,n)$ are MCLP coefficients predicting the reverberation at the output of the GSC. The filters $\mathbf{w}_{\mathrm{SC}}(n)$ and $\mathbf{w}_{\mathrm{LP}}(l,n)$ are estimated jointly by a Kalman filter [15] or using RLS [17] by considering $e_{\mathrm{ISCLP}}(n)$ as the error signal. In both methods, the steering vector $\mathbf{g}$ is modeled as the RETF, aiming at fully preserving early reflections.

4. MVDR BEAMFORMER WITH LINEAR PREDICTION

In this section, we propose a method for joint adaptive MVDR beamforming and MCLP based reverberation cancellation at the beamformer output. We refer to the proposed method as the *constrained Kalman beamformer with linear prediction* (CKBLP). In contrast to the GSC method, which employs a blocking matrix and minimizes the output signal power, we directly employ a constrained minimization without the need for a blocking matrix. Firstly, we define the adaptive beamformer output as

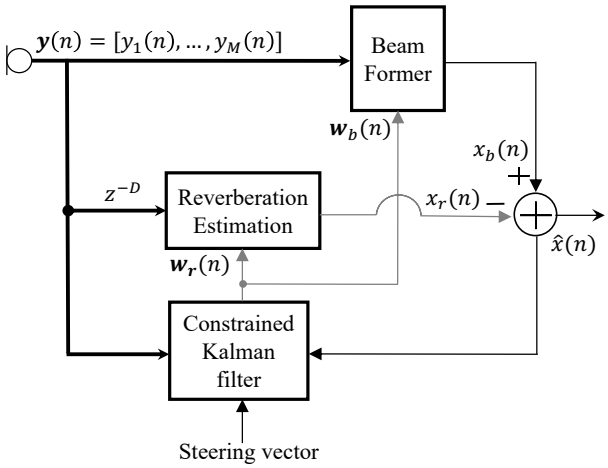

Fig. 1: The structure of joint beamforming and dereverberation using a constrained Kalman filter.

$$x_b(n) = \mathbf{w}_b^H(n)\mathbf{y}(n), \tag{6}$$

where $\mathbf{w}_b$ are the beamformer coefficients. Note that the beamformer branch $x_b(n)$ only filters the current frame. Secondly, using the MCLP model (3), the reverberation at the output of the beamformer is given by

$$\begin{aligned} x_r(n) &= \sum_{l=D}^{L} \mathbf{w}_b^H(n)\mathbf{W}_{\mathrm{r},l}(n)\mathbf{y}(n-l) \\ &= \sum_{l=D}^{L} \mathbf{w}_{\mathrm{r},l}^H(n)\mathbf{y}(n-l), \end{aligned} \tag{7}$$

where $\mathbf{w}_{\mathrm{r},l}(n)$ are the prediction coefficients after beamforming. As shown in Fig.1, the desired signal is then estimated by subtracting the predicted reverberation from the output of the beamformer as

$$\hat{x}(n) = x_b(n) - x_r(n). \tag{8}$$

4.1. Derivation of constrained Kalman filter

We derive a Kalman filter to jointly estimate the beamformer and reverberation prediction weights $\mathbf{w}_b(n)$ and $\mathbf{w}_{\mathrm{r},l}(n)$. To obtain a compact vector notation, we insert (6) and (7) into (8), i. e.,

$$\begin{aligned} \hat{x}(n) &= \mathbf{w}_b^H(n)\mathbf{y}(n) - \sum_{l=D}^{L} \mathbf{w}_{\mathrm{r},l}^H(n)\,\mathbf{y}(n-l) \\ &= \underbrace{\begin{bmatrix}\mathbf{w}_b^H(n) & -\mathbf{w}_{\mathrm{r},D}^H(n) & \cdots & -\mathbf{w}_{\mathrm{r},L}^H(n)\end{bmatrix}}_{\tilde{\mathbf{w}}^H(n)} \underbrace{\begin{bmatrix}\mathbf{y}(n) \\ \mathbf{y}(n-D) \\ \vdots \\ \mathbf{y}(n-L)\end{bmatrix}}_{\tilde{\mathbf{y}}(n)}, \end{aligned} \tag{9}$$

where $\tilde{\mathbf{y}}(n)$ are stacked microphone signals, and $\tilde{\mathbf{w}}(n)$ are stacked beamformer and reverberation prediction coefficient vectors, respectively [16, 17]. Now we can estimate the weight coefficients $\tilde{\mathbf{w}}(n)$ by minimizing the power of the desired signal $\hat{x}(n)$ under a distortionless constraint for the steering vector, i.e.,

$$\min_{\tilde{\mathbf{w}}} E[|\tilde{\mathbf{w}}^H(n)\tilde{\mathbf{y}}(n)|^2] \quad s.t. \quad \tilde{\mathbf{w}}^H\tilde{\mathbf{g}} = 1, \tag{10}$$

where $\tilde{\mathbf{g}} = [\mathbf{g}^T, 0, 0, \cdots, 0]^T$ is a vector containing $\mathbf{g}$ and zero-padded by $M(L-D+1)$ zeros. We can reformulate the observation and the constraint into a two-equation system [6, 24], where the first row is obtained by re-arranging the complex conjugate of (9), and the second row is the distortionless constraint in (10). Then, the measurement equation [6, 24] of the constrained Kalman filter is written as

$$\underbrace{\begin{bmatrix}\tilde{\mathbf{y}}^H(n)\\ \tilde{\mathbf{g}}^H\end{bmatrix}}_{\mathbf{P}(n)}\tilde{\mathbf{w}}(n) + \underbrace{\begin{bmatrix}-\hat{x}^*(n)\\ \epsilon_g(n)\end{bmatrix}}_{\boldsymbol{\epsilon}(n)} = \underbrace{\begin{bmatrix}0\\1\end{bmatrix}}_{\mathbf{c}}, \tag{11}$$

where $\epsilon_g(n)$ is the steering error, modeling inaccuracies between the estimated and true steering vector, which we model independently from $\hat{x}(n)$ as a zero-mean normal random variable, and * denotes the complex conjugate. $\boldsymbol{\epsilon}(n) = [-\hat{x}^*(n)\ \ \epsilon_g(n)]^T$ is the measurement noise vector with the correlation matrix

$$\boldsymbol{\Phi}_\epsilon(n) = \begin{bmatrix}\Phi_x(n) & 0\\ 0 & \Phi_g(n)\end{bmatrix}. \tag{12}$$

The unknown evolution of the time-varying filter $\tilde{\mathbf{w}}(n)$ can be modeled by a state-vector using a first-order Markov process

$$\tilde{\mathbf{w}}(n) = \mathbf{A}\tilde{\mathbf{w}}(n-1) + \mathbf{v}_w(n), \tag{13}$$

where the matrix $\mathbf{A}$ is a prediction matrix and $\mathbf{v}_w(n)$ is the driving-noise vector modeled by a zero-mean Gaussian random process with the covariance matrix $\boldsymbol{\Phi}_v(n)$.

From the observation and state equations (11) and (13), we estimate the weight vector coefficients recursively using the Kalman filter [25] by

$$\tilde{\mathbf{w}}(n|n-1)) = \mathbf{A}\tilde{\mathbf{w}}(n-1|n-1) \tag{14}$$

$$\mathbf{M}(n|n-1) = \mathbf{A}\mathbf{M}(n-1|n-1)\mathbf{A}^H + \boldsymbol{\Phi}_\mathbf{v}(n) \tag{15}$$

$$\mathbf{k}(n) = \mathbf{M}(n|n-1)\mathbf{P}^H(n) \times \Big(\boldsymbol{\Phi}_\epsilon(n) + \mathbf{P}(n)\mathbf{M}(n|n-1)\mathbf{P}^H(n)\Big)^{-1} \tag{16}$$

$$\tilde{\mathbf{w}}(n|n) = \tilde{\mathbf{w}}(n|n-1) + \mathbf{k}(n)(\mathbf{c} - \mathbf{P}(n)\tilde{\mathbf{w}}(n|n-1)) \tag{17}$$

$$\mathbf{M}(n|n) = \mathbf{M}(n|n-1) - \mathbf{k}(n)\mathbf{P}(n)\mathbf{M}(n|n-1). \tag{18}$$

Here, $\mathbf{M}(n)$ is the $M(L-D+2) \times M(L-D+2)$ estimation error covariance matrix, and $\mathbf{k}(n)$ is the Kalman filter gain.

4.2. Parameter estimation

We propose to model the estimation error covariance matrix as a diagonal matrix with fixed errors corresponding to the temporal update variances of the beamformer and linear prediction coefficients, Φ_b and Φ_p, respectively. The filter error covariance matrix is then given by

$$\boldsymbol{\Phi}_\mathbf{v}(n) = \mathrm{diag}\left\{[\underbrace{\Phi_b, \ldots, \Phi_b}_{M}, \underbrace{\Phi_p, \ldots, \Phi_p}_{M(L-D+1)}]^T\right\}, \tag{19}$$

where $\mathrm{diag}\{\}$ constructs a matrix with the argument on the main diagonal and zero elsewhere.

The desired signal PSD can be estimated using the decision-directed approach [2, 26] as a weighting between the current estimate using the previously estimated filter coefficients and the actual estimate of the previous frame, i. e.,

$$\Phi_x(n) = \beta|\hat{x}(n-1)|^2 + (1-\beta)|\tilde{\mathbf{w}}^H(n-1)\tilde{\mathbf{y}}(n)|^2, \tag{20}$$

where $0 < \beta < 1$ is the decision-directed weighting factor.

The steering vector $\mathbf{g}(n)$ can be estimated either as the RETF as in prior work [15, 16], or as the DRTF. The advantage of using DRTFs is that the beamformer will also reduce early reflections to some extent in contrast to using RETFs which fully preserves the early reflections. The DRTF steering vector can be estimated e. g. using the recently proposed approach in [27] based on spatial probabilities, modelling the DRTFs assuming ideal omni-directional microphones from the array geometry. The RETF steering can be estimated using the generalized eigendecomposition method described in [15]. To make the implementation online capable, the non-causal averaging operation of the RETFs over the whole audio file can be replaced by causal recursive averaging similarly as in [28], equations (39) and (40).

5. EXPERIMENTAL RESULTS

In this section, the CKBLP and the ISCLP methods are evaluated using the REVERB challenge dataset [21] which gives insights on the overall system performance compared to WPD-RLS method [17]. We present the results considering both steering vector estimation approaches based on the DRTFs and RETFs, respectively.

5.1. Experimental setup

We utilize the original evaluation metrics from the REVERB challenge [21, 29], i. e., perceptual evaluation of speech quality (PESQ), cepstral distance (CD), frequency-weighted segmental SNR (fwsSNR), speech-to-reverberation modulation energy ratio (SRMR) and log likelihood ratio (LLR) [29]. We used a subset of 200 files of the REVERB simulated development dataset (*SimData_dt*) for optimizing parameters and show results for the whole evaluation set. The evaluation data contains acoustic conditions in three different rooms with reverberation time (T60) of about 0.25 s, 0.5 s, 0.7 s, and two speaker-microphone distances, 50 cm (near), and 250 cm (far). An 8-channel circular array with 20 cm diameter was used. The simulated evaluation dataset contains recorded background noise with a signal-to-noise ratio (SNR) of about 20 dB.

In our experiments, we used a sampling rate of 16 kHz, and a 512 point STFT with 50% overlap. The prediction delay was $D = 2$ frames and the MCLP filter length varied for each frequency band with $L = \{12, 15, 6\}$ from low to high with transition frequencies $\{800, 2000\}$ Hz. We initialized $\mathbf{M}(0|0) = \boldsymbol{\Phi}_\mathbf{v}(0)$ using $\Phi_b(0) = 10^{-5}$, $\Phi_p(0) = 0.03$, and chose $\Phi_b(n) = 4 \times 10^{-7}$, $\Phi_p(n) = 6\times10^{-6}$ and $\Phi_g(n) = 10^{-12}$. We chose $\mathbf{A} = \mathbf{I}_{M\times(L-D+2)}$, where $\mathbf{I}$ is the identity matrix. All parameters were obtained by tuning on the development subset.

5.2. Results

Tables 1 and 2 show the REVERB challenge simulated data evaluation set results for each room and distance condition, and the average results. We show the unprocessed direct sound, and two variants of each of the proposed CKBLP and ISCLP methods, using either

Table 1: Objective speech enhancement evaluation using REVERB challenge dataset for different rooms and microphone distances (near and far). The steering vector is estimated using the DRTF method. Boldface shows the best performance.

Unprocessed					
Condition	PESQ	CD	fwsSNR	SRMR	LLR
Room 1 near	3.02	1.99	8.13	4.50	0.35
Room 1 far	2.28	2.67	6.68	4.58	0.38
Room 2 near	2.04	4.63	3.35	3.74	0.49
Room 2 far	1.66	5.21	1.04	2.97	0.75
Room 3 near	1.92	4.38	2.27	3.57	0.65
Room 3 far	1.57	4.96	0.24	2.73	0.84
Average	2.08	3.98	3.62	3.68	0.57
ISCLP-DRTF					
Room 1 near	1.96	4.60	2.23	4.78	0.64
Room 1 far	1.64	4.82	1.11	3.51	0.70
Room 2 near	1.96	4.69	-0.75	2.77	1.00
Room 2 far	1.66	5.28	-0.86	2.65	1.00
Room 3 near	1.97	4.76	1.64	3.75	0.82
Room 3 far	1.78	5.01	0.69	2.98	0.94
Average	1.83	4.86	0.68	3.41	0.85
CKBLP-DRTF (Proposed)					
Room 1 near	3.31	1.88	10.06	4.88	0.34
Room 1 far	2.67	2.22	9.04	5.28	0.39
Room 2 near	2.47	3.88	6.72	4.58	0.40
Room 2 far	2.04	4.42	4.93	4.80	0.52
Room 3 near	2.36	3.53	5.37	4.61	0.52
Room 3 far	1.94	3.97	3.89	4.42	0.59
Average	**2.46**	**3.32**	**6.67**	**4.76**	**0.46**

Table 2: Objective speech enhancement evaluation using REVERB challenge dataset for different rooms and microphone distances (near and far). The steering vector is estimated using the RETF method. Boldface shows the best performance.

Condition	PESQ	CD	fwsSNR	SRMR	LLR
ISCLP-RETF					
Room 1 near	2.87	2.37	7.64	4.30	0.43
Room 1 far	2.35	2.97	6.43	4.67	0.48
Room 2 near	2.30	4.20	4.90	4.04	0.45
Room 2 far	1.89	4.85	3.30	4.12	0.62
Room 3 near	2.26	3.92	3.72	4.07	0.63
Room 3 far	1.82	4.49	2.05	3.75	0.75
Average	2.25	3.80	4.67	4.16	0.56
CKBLP-RETF (Proposed)					
Room 1 near	2.41	3.10	6.68	4.83	0.45
Room 1 far	2.83	2.52	7.90	4.80	0.42
Room 2 near	2.06	4.27	4.95	4.79	0.53
Room 2 far	2.66	3.23	7.41	4.57	0.41
Room 3 near	1.94	4.03	4.12	4.56	0.57
Room 3 far	2.35	3.80	5.52	4.52	0.48
Average	**2.37**	**3.49**	**6.10**	**4.68**	**0.47**

the DRTF steering vector [27], referred to as the CKBLP-DRTF and ISCLP-DRTF, or using the RETF steering vector [15], referred to as the CKBLP-RETF and ISCLP-RETF.

As shown in Table 1, the proposed CKBLP-DRTF method achieves the best performance for the average of the results in all types of room conditions and also for each room condition individually. While we confirmed the ISCLP-DRTF improves STOI [30], it here degrades most of our considered metrics in most conditions. The GSC is known to suffer from speech cancellation, which is prevented by using RETFs. In contrast, the CKBLP performs well using both RETFs and DRTFs. As shown in Table 2, the CKBLP-RETF outperforms the ISCLP-RETF method for the average of all conditions. In addition, the CKBLP-DRTF also slightly outperforms the online system proposed in [17] with single pass over the data (i.e., true online processing), which uses a neural network based RETF estimator on pre-dereverberated signals using an additional weighted prediction error (WPE) pre-processing stage. In contrast, our proposed CKBLP-DRTF method is a more direct and potentially less complex system, as it only uses a low-complexity DOA estimator and then directly estimates the beamformer. Using the proposed CKBLP-DRTF, the SRMR is enhanced by 30% compared to the unprocessed signals, 15% compared to the ISCLP-RETF method, and 40% compared to the ISCLP-DRTF method. Also, the averaged fwsSNR is improved by 84% compared to the unprocessed signals, and 43% compared to the ISCLP-RETF method.

Employing the spatial probability-based DRTFs in the proposed method not only does improve the performance of the speech enhancement systems, but also imposes lower computational complexity to the system. Particularly, the beamformer using the DRTFs partially reduces early reflections in contrast to the RETFs which maintain early reflections. Although it is difficult to exactly compare the computational complexity of steering vector estimation methods, we measured the total runtime of our non-optimized Matlab implementations. The total runtime using the DRTF-based steering vector was about 30% faster than the eigenvalue decomposition based RETF steering vector.

6. CONCLUSIONS

In this paper, we have presented a new algorithm for joint beamforming and reverberation cancellation of speech signals. The proposed algorithm, which is derived based on the constrained Kalman filter with multichannel linear prediction, jointly estimates the beamformer and reverberation canceler filters. Experimental results for the REVERB challenge dataset show that the proposed method significantly reduces the noise and reverberation, and improves the speech quality compared to state-of-the-art methods. While existing GSC-based integrated solution degrades the speech when using DRTF-based steering vector, our proposed method achieves significant improvements using both the RETF- and DRTF-based steering vectors, where the latter yields a better performance while being simpler to compute.

7. REFERENCES

[1] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[2] S. Braun and E. A. Habets, "Linear prediction-based online dereverberation and noise reduction using alternating Kalman filters," *IEEE/ACM Trans. on Audio, Speech and Language Proc.*, vol. 26, no. 6, pp. 1115–1125, 2018.

[3] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE AAS magazine*, vol. 5, no. 2, pp. 4–24, 1988.

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

[5] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 17, no. 6, pp. 1071–1086, 2009.

[6] Y.-H. Chen and C.-T. Chiang, "Adaptive beamforming using the constrained Kalman filter," *IEEE Transactions on Antennas and Propagation*, vol. 41, no. 11, pp. 1576–1580, 1993.

[7] A. El-Keyi, T. Kirubarajan, and A. B. Gershman, "Robust adaptive beamforming based on the Kalman filter," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 3032–3041, Aug 2005.

[8] O. Schwartz, S. Gannot, and E. A. P. Habets, "Nested generalized sidelobe canceller for joint dereverberation and noise reduction," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 106–110.

[9] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 20, no. 10, pp. 2707–2720, 2012.

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

[11] M. Togami, "Multichannel online speech dereverberation under noisy environments," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, Aug 2015, pp. 1078–1082.

[12] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, S. Araki, T. Hori, et al., "Strategies for distant speech recognitionin reverberant environments," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 60, 2015.

[13] W. Yang, G. Huang, W. Zhang, J. Chen, and J. Benesty, "Dereverberation with differential microphone arrays and the weighted-prediction-error method," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 376–380.

[14] L. Drude, C. Boeddeker, J. Heymann, R. Haeb-Umbach, K. Kinoshita, M. Delcroix, and T. Nakatani, "Integrating neural network based beamforming and weighted prediction error dereverberation.," in *Interspeech*, pp. 3043–3047.

[15] T. Dietzen, S. Doclo, M. Moonen, and T. Van Waterschoot, "Joint multi-microphone speech dereverberation and noise reduction using integrated sidelobe cancellation and linear prediction," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sep. 2018, pp. 221–225.

[16] T. Nakatani and K. Kinoshita, "A unified convolutional beamformer for simultaneous denoising and dereverberation," *IEEE Signal Proc. Lett.*, vol. 26, no. 6, pp. 903–907, 2019.

[17] T. Nakatani and K. Kinoshita, "Simultaneous denoising and dereverberation for low-latency applications using frame-by-frame online unified convolutional beamformer," in *INTERSPEECH 2019*, 2019.

[18] I. Kodrasi and S. Doclo, "Late reverberant power spectral density estimation based on an eigenvalue decomposition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 611–615.

[19] C. Boeddeker, H. Erdogan, T. Yoshioka, and R. Haeb-Umbach, "Exploring practical aspects of neural mask-based beamforming for far-field speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6697–6701.

[20] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 196–200.

[21] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, et al., "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.

[22] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, et al., "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 7, 2016.

[23] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.

[24] D. Cherkassky and S. Gannot, "New insights into the Kalman filter beamformer: Applications to speech and robustness," *IEEE Signal Proc. Lett.*, vol. 23, no. 3, pp. 376–380, 2016.

[25] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. of the ASME Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[26] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, December 1984.

[27] S. Braun and I. Tashev, "Acoustic localization using spatial probability in noisy and reverberant environments," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2019, pp. 1–5.

[28] O. Schwartz, S. Gannot, and E. A. Habets, "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions," *IEEE/ACM Trans. on Audio, Speech, and Language Proc.*, vol. 23, no. 2, pp. 240–251, 2014.

[29] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 16, no. 1, pp. 229–238, 2007.

[30] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE/ACM Trans. on Audio, Speech, and Language Proc.*, vol. 19, no. 7, pp. 2125–2136, Sept 2011.