MULTICHANNEL ACOUSTIC ECHO CANCELATION IN MULTIPARTY SPATIAL AUDIO CONFERENING WITH CONSTRAINED KALMAN FILTERING

Zhengyou Zhang, Qin Cai, Jack W. Stokes

Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA

ABSTRACT

This paper proposes a novel scheme for acoustic echo cancellation (AEC) in multi-party, spatialized audio conferencing by exploring the constraints among the adaptive filters for each synthesized remote speaker. The AEC algorithm employs constrained Kalman filtering (CKF-AEC) that takes advantage of the uncorrelated reference signal from each remote channel [1], yet restricts the filter adaptation within the subspace determined by the spatialization functions used to virtualize each remote participant and the number of unique acoustic paths between the loudspeakers and microphone. In addition, the proposed algorithm allows the adaptation for channels without input signals. Experimental results show the proposed algorithm has much higher performance than the scheme proposed in [1] which uses NLMS as the adaptive filter and does not explore the available constraint.

Index Terms— Spatial Audio, Multi-channel AEC, Kalman Filtering, Multi-party conferencing.

INTRODUCTION

Recently, the demand for better quality teleconferencing with remote multi-participants has risen rapidly for reducing travel cost and increasing productivity. It has been shown that spatial mapping of the remote participants' voice signals in multiple virtual positions enhances the collaboration experience during multi-party conferencing [2]. It is well known that stereo AEC suffers from the mis-convergence problem [3]. Researchers have proposed various ways to alleviate the problem including adding nonlinearities [4] or additional noise [3,5] to de-correlate the speaker signals. In the paper, we propose an acoustic echo cancellation algorithm based on the constrained Kalman filter (CKF-AEC) for full-duplex, multi-party, spatialized audio conferencing using Voice over Internet Protocol (VoIP) which does not distort the speaker signal. Kalman filtering has not been used widely for AEC except in [6] which directly considers the varying acoustic paths as the hidden state variable. The multi-party conferencing problem with spatialized audio was previously studied in [1], and we adopt the same synthetic stereophonic structure depicted in Fig. 1. However in [1], each canceller is adapted independently without realizing that the only unknown parameters are the room impulse responses (RIRs) between the loudspeakers and microphones; the method also ignores the fact that signals are continuously sent to each loudspeaker. If one remote participant has been silent for a long time, noticeable echo will be heard when the participant resumes speaking. Another disadvantage of the method in [1] is that it has very poor performance when multiple remote participants speak simultaneously; the authors in [1] analyzed the convergence for adapting four remote channels talking at the same time using white Gaussian noise. However in our experiments with speech signals, the joint NLMS-based multi-channel AEC failed.

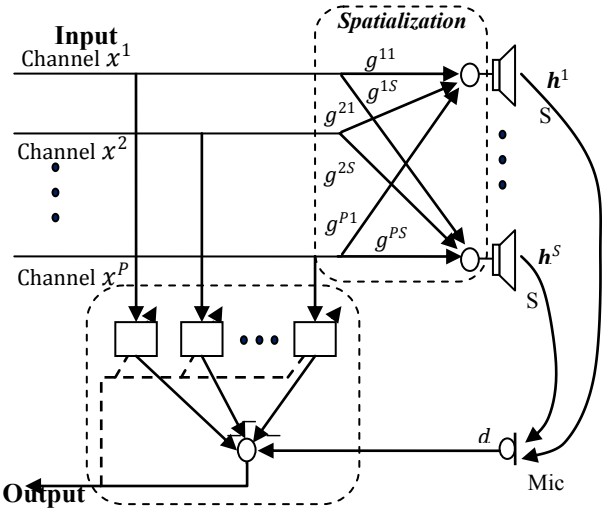

Fig. 1 Synthetic stereo structure with a single canceller per remote-channel

The CKF-AEC algorithm proposed in this paper employs a constrained Kalman filtering mechanism using the signal from each remote channel as reference. It takes advantage that the signals from each remote channel are uncorrelated and the spatialization parameters are known. Not only does the algorithm adapt the per channel impulse response (CIR), it also resolves the RIR between each loudspeaker and microphone. As long as one channel is active, the RIRs and CIRs for the other channels are adapted simultaneously. In other words, if a remote participant is silent for a long time, the adaptation from the other channels will benefit the

adaptation for this participant. The CKF-AEC algorithm does not require channel switching as in multiple mono NLMS-based AEC; it even adapts with multiple remote participants speaking simultaneously. Finally to significantly reduce the CPU consumption, we propose an implementation to avoid the matrix inversion that is $O(n^3)$ for the standard KF.

CONSTRAINED KALMAN FILTERING

We describe the proposed algorithm in the frequency domain due to ease of the implementation since the spatial mapping only results in multiplication of a complex number in the frequency domain. The following derivation is conducted in each frequency sub-band.

2.1. Problem Statement

The superscript * denotes the complex conjugate. Column vectors are assumed. The superscript ^{T} is the matrix/vector transpose. The superscript ^{H} is the Hermitian transpose (complex conjugate of the transpose). The subscript $_t$ will be used to indicate time instant (or frame number). Assume there are P remote participants/channels with audio inputs $\{X^i | i = 1, \dots, P\}$ and S loudspeakers. Each remote audio channel X^i is spatialized across the loudspeakers with gain and delay modulation G^{is} for speaker s. The audio played on speaker s is thus $Y^s = \sum_{i=1}^{P} G^{is} X^i$. A microphone is used to capture the audio in the local room, and is denoted by D. The captured audio includes the audio of the local participant and that played on the speakers. The audio played on speaker s is transmitted through the RIR modeled by an FIR filter with L taps, denoted by $\boldsymbol{H}_t^s = [H_t^s, H_{t-1}^s, \dots, H_{t-L-1}^s]^T$.

Assuming that the local or near end participant is not talking, the audio captured by the microphone at time t, i.e., the echo, should be

$$D_t = \sum_{i=1}^{P}\sum_{s=1}^{S} G^{is}\left(H_t^s X_t^i + H_{t-1}^s X_{t-1}^i + \cdots + H_{t-L-1}^s X_{t-L-1}^i\right) = \sum_{i=1}^{P}\sum_{s=1}^{S} G^{is}\, \boldsymbol{H}_t^{s\,T} \boldsymbol{X}_t^i$$

where $\boldsymbol{X}_t^i = \left[X_t^i, X_{t-1}^i, \dots, X_{t-L-1}^i\right]^T$ is speech from remote participant i.

The problem is to design an echo cancellation filter with L taps for each remote participant i, $\boldsymbol{W}_t^i = \left[W_t^i, W_{t-1}^i, \dots, W_{t-L-1}^i\right]^T$ such that the echo is cancelled. That is, determine $\boldsymbol{W}_t^i$'s so

$$D_t - \sum_{i=1}^{P} \boldsymbol{W}_t^{i\,T} \boldsymbol{X}_t^i = 0. \qquad (1)$$

It is clear that we have

$$\boldsymbol{W}_t^i = \sum_{s=1}^{S} G^{is} \boldsymbol{H}_t^s. \qquad (2)$$

Therefore, the echo cancellation filters are not mutually independent. We will leverage this constraint to update each echo cancellation filter simultaneously even though the corresponding remote participant is not talking. If the local participant is talking, then the system outputs the echo-cancelled signal: $D_t - \sum_{i=1}^{P} \boldsymbol{W}_t^{i\,T} \boldsymbol{X}_t^i$.

2.2. Kalman Filtering Formulation

Let the state vector be a $(P+S)L$-dimensional vector:

$$\boldsymbol{S}_t = \left[\boldsymbol{W}_t^{1\,T}, \dots, \boldsymbol{W}_t^{P\,T}, \boldsymbol{H}_t^{1\,T}, \dots, \boldsymbol{H}_t^{S\,T}\right]^T = [W_t^1, W_{t-1}^1, \dots, W_{t-L-1}^1, \dots, W_t^P, W_{t-1}^P, \dots, W_{t-L-1}^P, H_t^1, H_{t-1}^1, \dots, H_{t-L-1}^1, \dots, H_t^S, H_{t-1}^S, \dots, H_{t-L-1}^S]^T.$$

The state is not expected to remain constant, and the state equation is

$$\boldsymbol{S}_t = \boldsymbol{S}_{t-1} + \boldsymbol{n}_t \qquad (3)$$

where $\boldsymbol{n}_t$ is the system noise. The parts of $\boldsymbol{n}_t$ corresponding to the H elements model the variation due to changes in the acoustical environment such as the movement of the local participant. The observation equation is:

$$D_t = \boldsymbol{A}_t^T \boldsymbol{S}_t + v_t, \qquad (4)$$

where v_t is the observation noise (microphone noise and ambient noise) and can also model the inaccuracy of the observation system. The vector $\boldsymbol{A}_t$ is a $(P+S)L$-dimensional vector given by

$$\boldsymbol{A}_t = \left[\boldsymbol{X}_t^{1\,T}, \dots, \boldsymbol{X}_t^{P\,T}, \boldsymbol{0}^{1\,T}, \dots, \boldsymbol{0}^{S\,T}\right]^T$$

with $\boldsymbol{0}$ being the L-dimensional zero vector. The constraint on the state parameters (2) is a set of PL linear equations, and can be written as

$$\boldsymbol{C}\boldsymbol{S}_t = \boldsymbol{0} \qquad (5)$$

where $\boldsymbol{C}$ is a $PL \times (P+S)L$ matrix, given by

$$\boldsymbol{C} = \begin{bmatrix} -\boldsymbol{1} & & G^{11}\boldsymbol{1} & \cdots & G^{1S}\boldsymbol{1} \\ & \ddots & \vdots & \ddots & \vdots \\ & -\boldsymbol{1} & G^{P1}\boldsymbol{1} & \cdots & G^{PS}\boldsymbol{1} \end{bmatrix}$$

with $\boldsymbol{1}$ being the $L \times L$ identity matrix, i.e., $\boldsymbol{1} = \mathrm{diag}(1, \dots, 1)$. The state constraint can be considered as perfect observation. Thus, we combine the original observation equation and the state constraint into a new observation equation:

$$\boldsymbol{Y}_t = \boldsymbol{B}_t \boldsymbol{S}_t + \boldsymbol{v}_t \qquad (6)$$

where

$$\boldsymbol{Y}_t = [D_t, 0, \dots, 0]^T, \quad \boldsymbol{B}_t = \begin{bmatrix} \boldsymbol{A}_t^T \\ \boldsymbol{C} \end{bmatrix}, \quad \text{and} \quad \boldsymbol{v}_t = \begin{bmatrix} v_t \\ \boldsymbol{u}_t \end{bmatrix}$$

with $\boldsymbol{u}_t$ being the noise term for the state constraint. If the constraint is exact, then $\boldsymbol{u}_t = \boldsymbol{0}$.

We assume that the following conditions are satisfied: $E[\boldsymbol{n}_t] = \boldsymbol{0}$, $E[\boldsymbol{v}_t] = \boldsymbol{0}$, $E[\boldsymbol{n}_t \boldsymbol{n}_t^T] = \boldsymbol{Q}_t$ and $E[\boldsymbol{v}_t \boldsymbol{v}_t^T] = \boldsymbol{R}_t = \begin{bmatrix} \sigma_t^2 & \boldsymbol{0}^T \\ \boldsymbol{0} & \boldsymbol{\Lambda}_t \end{bmatrix}$

where $\boldsymbol{\Lambda}_t$ is the covariance matrix of the noise term for the state constraint. If we indeed want to impose the constraint fully, then $\boldsymbol{\Lambda}_t = \mathbf{0}$. In practice, we sometimes prefer to impose a soft constraint to have a more stable system on account of nonlinearity in loudspeakers and clock drift between loudspeaker and sound capture. We can also start $\boldsymbol{\Lambda}_t$ with a larger value and gradually decrease it over time.

We can now solve the multichannel AEC problem for spatialized audio using the Kalman filter. We use superscript "$^-$" for the prediction, and $\boldsymbol{P}$ for the covariance matrix of the error in the estimated state vector. The Kalman filter equations are given by:

$$\begin{aligned}
\boldsymbol{S}_t^- &= \boldsymbol{S}_{t-1} \\
\boldsymbol{P}_t^- &= \boldsymbol{P}_{t-1} + \boldsymbol{Q}_t \\
\boldsymbol{K}_t &= \boldsymbol{P}_t^- \boldsymbol{B}_t^H (\boldsymbol{B}_t \boldsymbol{P}_t^- \boldsymbol{B}_t^H + \boldsymbol{R}_t)^{-1} \\
\boldsymbol{S}_t &= \boldsymbol{S}_t^- + \boldsymbol{K}_t (\boldsymbol{Y}_t - \boldsymbol{B}_t \boldsymbol{S}_t^-) \\
\boldsymbol{P}_t &= (\mathbf{I} - \boldsymbol{K}_t \boldsymbol{B}_t) \boldsymbol{P}_t^-
\end{aligned}$$

The state vector can be initialized to 0, i.e., $\boldsymbol{S}_0^- = \mathbf{0}$. The covariance matrix $\boldsymbol{P}_0^-$ should be set to a large value to reflect that we do not have knowledge of the state vector.

The advantages of the CKF algorithm are: 1) The constraint is taken care of automatically, and can be imposed with varying degrees, 2) All channels are taken into account simultaneously. Thus, overlapping far-end talking is not an issue, 3) The AEC for each channel is updated continuously because of the constraint, even if that channel is not active. Therefore, AEC is always up to date. 4) Ambient noise σ_t^2 can be time varying. A separate noise tracker can be used to provide that information.

However, we are also aware of the major drawback is the need to invert a $(PL+1) \times (PL+1)$ matrix, rather than a scalar inversion. Matrix inversion has $O((PL+1)^3)$ complexity. Thus there is a significant increase in computational cost.

2.3. An Improved Implementation

In our previous discussion, we added the PL constraints to the set of measurement equations, and the constraints are imposed softly by a Gaussian noise vector $\boldsymbol{u}_t$ with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Lambda}_t$. In practice, the covariance matrix $\boldsymbol{\Lambda}_t$ is usually set to be a diagonal matrix by assuming no correlation between the constraints. Thus the covariance matrix of the measurement noise vector $\boldsymbol{R}_t$ is diagonal.

Let $v_{t,j}$ be the j-th element of the measurement noise vector $\boldsymbol{v}_t$, $\boldsymbol{R}_t = \mathrm{diag}(r_{t,1}, \dots, r_{t,j}, \dots, r_{t,PL})$ be the diagonal covariance matrix, $\boldsymbol{b}_{t,j}^T$ be the j-th row of the measurement matrix $\boldsymbol{B}_t$, and $Y_{t,j}$ be the j-th element of the measurement vector $\boldsymbol{Y}_t$, then the vector measurement equation (6) is equivalent to the following $PL+1$ scalar measurement equations:

$$Y_{t,j} = \boldsymbol{b}_{t,j}^T \boldsymbol{S}_t + v_{t,j} \quad \text{for } j = 1, \dots, PL+1$$

with $E[v_{t,j}] = 0$ and $E[v_{t,j}^2] = r_{t,j}$. Now, we can apply the Kalman filter sequentially for each of the above scalar measurements, and inversion of a $(PL+1) \times (PL+1)$ matrix is avoided. The complexity is reduced from $O((PL+1)^3)$ to $O(PL+1)$. This substantially reduces the computational cost.

In summary, the new algorithm can be described as follows:

Prediction:

$$\begin{aligned}
\boldsymbol{S}_t^- &= \boldsymbol{S}_{t-1} \\
\boldsymbol{P}_t^- &= \boldsymbol{P}_{t-1} + \boldsymbol{Q}_t
\end{aligned}$$

Update: Let $\boldsymbol{S}_{t,0}^- \triangleq \boldsymbol{S}_t^-$, and $\boldsymbol{P}_{t,0}^- \triangleq \boldsymbol{P}_t^-$.

For $j = 1, \dots, PL+1$, do

$$\begin{aligned}
\boldsymbol{K}_{t,j} &= \boldsymbol{P}_{t,j-1}^- \boldsymbol{b}_{t,j}^* \left(\boldsymbol{b}_{t,j}^T \boldsymbol{P}_{t,j-1}^- \boldsymbol{b}_{t,j}^* + r_{t,j} \right)^{-1} \\
\boldsymbol{S}_{t,j}^- &= \boldsymbol{S}_t^- + \boldsymbol{K}_{t,j} (Y_{t,j} - \boldsymbol{b}_{t,j}^T \boldsymbol{S}_{t,j-1}^-) \\
\boldsymbol{P}_{t,j}^- &= \left(\mathbf{I} - \boldsymbol{K}_{t,j} \boldsymbol{b}_{t,j}^T \right) \boldsymbol{P}_{t,j-1}^-
\end{aligned}$$

Finally, we set $\boldsymbol{S}_t \triangleq \boldsymbol{S}_{t,PL+1}^-$ and $\boldsymbol{P}_t \triangleq \boldsymbol{P}_{t,PL+1}^-$.

If a constraint is already satisfied, then $Y_{t,j} - \boldsymbol{b}_{t,j}^T \boldsymbol{S}_{t,j-1}^- = 0$, leaving the state vector unaffected. If not, a correction is added by projecting to the direction $\boldsymbol{K}_{t,j} \propto \boldsymbol{P}_{t,j-1}^- \boldsymbol{b}_{t,j}^*$.

EXPERIMENTAL RESULTS

We consider four remote participants with spatialization on two loudspeakers with virtual positions at [-30˚, 30˚, 0˚, -45˚], which is the angle between the virtual speaker and the center line originating from the microphone, i.e., 0˚ gives an illusion the participant is positioned in the middle front. We used 16kHz speech. Each remote participant speaks for 4 seconds. A known RIR is used to generate the microphone input and -20dB white Gaussian noise is added. The number of sub-bands is 512 with 256 samples per frame. The best result is achieved with σ_t = -20dB at the lowest bands, $\boldsymbol{\Lambda}_t$ = 1e-2 and $\boldsymbol{Q}_t = 0$. Fig. 2 shows the comparison between our CKF-AEC and the four mono NLMS filters proposed in [1] (but with our frequency domain implementation). Whenever a new talker starts, the Error Return Loss Enhancement (ERLE) drops significantly in NLMS. However after the second talker, thanks to the constraint, the CKF-AEC handles new participants gracefully. In Fig. 3, we conducted another experiment with two participants talking simultaneously in consecutive positions and compared the CKF-AEC algorithm with the joint NLMS which adapts parameters for four channels simultaneously, described in [1]. It is shown that joint NLMS performs poorly while the CKF-AEC maintains good performance. The authors in [1] only showed convergence with white Gaussian noise input, but not with speech signals. Finally, we add -30dB change on RIRs for every 50 frames and add $\boldsymbol{Q}_t$ accordingly. Fig 4 shows the case where the first channel is re-activated at the end of the third channel or at 12 seconds. This is to check whether CKF-AEC can update a canceller even when the

corresponding channel is silent. Indeed, CKF-AEC handles the transition more smoothly than the four mono NLMS after the second participant. We repeat the same experiment with Kalman filtering without the constraints, glitches similar to NLMS happened for each transition, as shown in Fig 5.

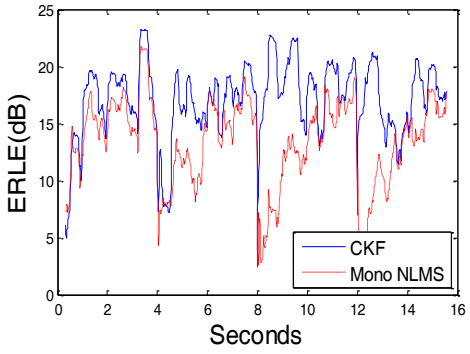

Fig. 2 AEC Performance between CKF and 4 mono NLMS

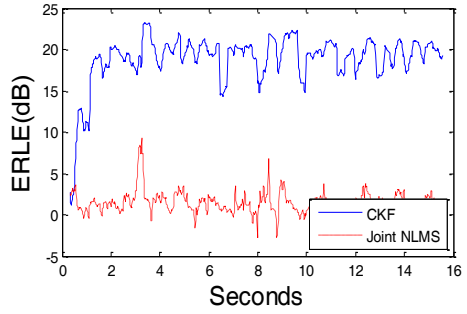

Fig. 3 AEC Performance between CKF and joint NLMS

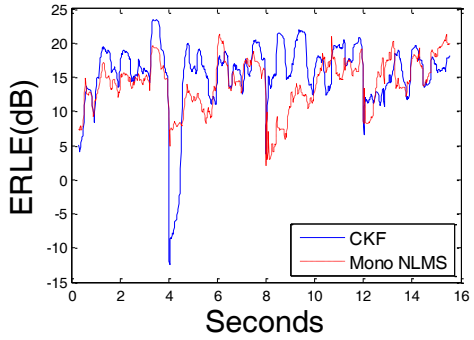

Fig. 4 AEC performance with changing RIRs

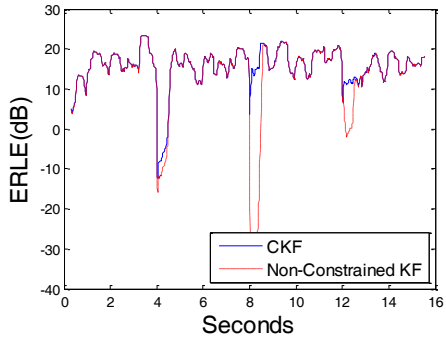

Fig 5. Performance between CKF and non-constrained KF

To verify if the adaptive filters converge to the optimal solution, we calculated the mean values of misalignments using known $\boldsymbol{S}_t$ in the frequency domain. Two data sets were used. Table 1 shows that CKF-AEC converges for both data sets. Mono NLMS also converges for each CIR, but without estimation of the RIRs.

	CKF	Mono NLMS
$\overline{misalign(w_i)}$ (set 1)	-22	-21
$\overline{misalign(h_i)}$ (set 1)	-19	N/A
$\overline{misalign(w_i)}$ (set 2)	-24	-18
$\overline{misalign(h_i)}$ (set 2)	-23	N/A

Table 1 Misalignment evaluation in dB.

CONCLUSION

We proposed a novel algorithm using Kalman filtering for multi-channel AEC during spatial audio conferencing by imposing constraints among each canceller during adaptation. The imposed constraints guide the adaptation toward the optimal solution. They allow for estimation of the RIRs as a byproduct and for adaptation for channels without input signals. Experimental results show that the algorithm outperforms the array of mono NLMS when a new channel starts. It also adapts when speech co-exists among multiple channels without need of channel switching. As our future work, we plan to compare our algorithm with stereo AEC algorithms such as the one proposed in [4].

ACKNOWLEDGMENTS

The authors thank Philip A. Chou, Dinei Florencio, and Zicheng Liu for many fruitful discussions.

REFERENCES

[1] T. N. Yensen, R. A. Goubran, and I. Lambadaris, "Synthetic Stereo Acoustic Echo Cancellation Structure for Multiple Participant VoIP Conference" *IEEE Transaction on Speech and Audio Processing*, vol. 9, no. 2, pp. 168-174, Feb. 2001.

[2] J. Baldis, "Effects of spatial audio on memory, compression, and preference during desktop conferences," *Proceedings of the CHI Conference on Human Factors in Computing Systems*, ACM Press, Seattle, 2001.

[3] M. Sonfhi, D. Morgan, and J. Hall, "Stereophonic acoustic echo cancelation – An overview of the fundamental problem," *IEEE Speech Processing Letter*, vol. 2, pp. 148-151, 1995.

[4] J. Benesty, D. Morgan, and M. Sonfhi, "A better understanding and an improved solution to the problem of stereophonic acoustic echo cancellation," *IEEE Transaction on Speech and Audio Processing*, vol. 6, pp. 156-165, 1998.

[5] A. Gilloire, and V. Turbin "Using auditory properties to improve the behavior of stereophonic acoustic cancellers," *Proceedings of IEEE Intl. Conf. Acoustic, Speech and Signal Processing*, pp. 3681-3684, 1998.

[6] G. Enzner, and P. Vary, "Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones," *Signal Processing*, vol. 86, pp. 1140-1156, Oct, 2006.