Clock Skew Robust Acoustic Echo Cancellation

Karim Helwani, Erfan Soltanmohammadi, Michael M. Goodwin, and Arvindh Krishnaswamy

Amazon Web Services, Inc., Palo Alto, CA, USA

{helwk, solterfa, mmg, arvindhk}@amazon.com

Abstract

Traditional acoustic echo cancelers require that the reference and microphone signals have exactly the same sampling frequency. In this paper, we present a novel Kalman filtering approach to acoustic echo cancellation (AEC) which blindly accounts for the clock skew between the playback and recording devices without the need for exchanging timestamps when they have independent clocks. The proposed Kalman-filter AEC employs a state-space formulation for the clock-skew problem and is implemented in a subband-domain based on our proposed novel variation of the complex modified discrete cosine transform which allows for arbitrary hop size and therefore enhanced time resolution. We show that the proposed algorithm is robust to Gaussian and super-Gaussian near-end noises and provide experimental results which demonstrate the state-of-the-art echo cancellation performance of the approach under clock-skew conditions.

Index Terms: Skew estimation, clock skew, multi-channel Kalman filter, subband filtering, robust adaptation

1. Introduction

In teleconferencing applications, there are disturbing echoes produced by the acoustic feedback of the loudspeaker signals into the microphones. Hence, an enabling technology for hands-free communication is acoustic echo cancellation (AEC), which aims at canceling the acoustic echoes from the microphone signals. In an AEC, the signal of the reproduction channel originates from speech or audio sources in a transmission room (far end). To cancel the echoes in the near end, an adaptation algorithm estimates a filter $\hat{\mathbf{h}}$, which models the acoustic path from the loudspeaker to the microphone. The loudspeaker signal x is then filtered with $\hat{\mathbf{h}}$ to generate a replica of the echo in the microphone signal. The resulting signal is then subtracted from the near-end microphone signal. If the estimated filter coefficients $\hat{\mathbf{h}}$ are equal to the true transfer path $\mathbf{h}$, then all echoes will be removed. For the echo cancellation problem, an underlying assumption is that the loudspeaker and microphone signals are synchronized. Typically, solutions either rely on synchronizing the signals using timestamps [1], or explicitly estimating the offset and resampling the signals [2–4]. Some other approaches are more tolerant to phase misalignment at the cost of quality, such as echo suppressors [5, 6]. Despite the superiority of deep learning solutions in suppressing the residual echoes, it has been shown that the employment of a high performance AEC improves the quality during a double-talk situations [7]. In this paper, we present a new class of AEC which is robust to clock skew without requiring timestamps for resynchronization.

2. Adaptive Blind Skew Estimation

The slight difference between the clock (sampling) frequencies of playback stream f_p and record stream f_r causes a performance drop in AEC. If we define $\delta = 1/f_r - 1/f_p$ and $\rho = 1/(1 + \delta f_p)$, then $f_r = \rho f_p$. If δ is nonzero, ρ is either greater than 1 or less than 1, which equivalently means the spectrum of the record signal is shrunk or expanded in comparison to the playback spectrum.

Consider the near end of communication system with loudspeaker signal $x(t)$, microphone signal $d(t)$, and transfer path response $h(t)$. The microphone signal can be modeled as:

$$d(t) = \int_{\tau} h(\tau)x(t-\tau)d\tau + \nu(t), \tag{1}$$

where $\nu(t)$ is the combined effect of noise and near-end signal. With clock skew, the discrete representation becomes

$$d(n/f_r) = \int_{\tau} h(\tau)x(n/f_r - \tau)d\tau + \nu(n/f_r), \tag{2}$$

where n is discrete time instance. Under clock skew, $x(n/f_r)$ is not available at the AEC. Instead, the AEC uses $x(n/f_p)$ along with $d(n/f_r)$ to estimate h. By the change of variable in (2), it can be shown that

$$d(n/f_r) = \int_{\tau'} h(n\delta + \tau')x(n/f_p - \tau')d\tau' + \nu(n/f_r). \tag{3}$$

In effect, the skew causes the channel seen from the microphone to expand or shrink over time. To account for the skew accurately, we require a subband transformation with high frequency and time resolution. To do so, we first introduce a new subband algorithm based on the complex modified discrete cosine transform.

2.1. Multi-hop Modified Complex Cosine Transform

The modified discrete cosine transform (MDCT) is often used in audio codecs due to low complexity, frequency selectivity and perfect reconstruction properties [8]. The complex version of the MDCT (CMDCT) allows linear filtering of a signal in the subband domain by minimizing the aliasing artifacts [9]. However, the original CMDCT has a fixed 50% overlap which limits the time resolution required for subband-domain AEC or other subband processing. Here we propose a multi-hop version of CMDCT called MH-CMDCT to work with arbitrary overlap size. First, we propose a tightening procedure on the analysis and synthesis windows so that they jointly act as a tight frame [10]. This procedure ensures that time-domain aliasing cancellation still works. The details of this are shown in Algorithm 1.

We modify the CMDCT and inverse CMDCT (ICMDCT) so they can run in a multi-hop manner using these windows. The details of the MH-CMDCT and the inverse MH-CMDCT (IMH-CMDCT) are shown in Algorithm 2 and Algorithm 3. Note that the FFT can be used to implement these algorithms efficiently. In the proposed AEC, the time-domain signals are transformed into subband-domain signals using the MH-CMDCT. From now on, all signals are in the subband-domain.

Algorithm 1 MH-CMDCT Window Tightening Procedure

```
Input: Initial analysis window $\psi_a(n)$, initial synthesis window $\psi_s(n)$, window size $N_w$, and hop size $N_h$
Output: Tight windows: $w_a(n)$ and $w_s(n)$
for $n \in \{0, 1, \dots N_w - 1\}$ do
    $\tilde{\psi}(n) = \psi_a(n)\psi_s(n)$
end for
$\tilde{\psi}(n) = 0$ for $n < 0$ and $n \geq N_w$
for $n \in \{0, 1, \dots N_w - 1\}$ do
    $\varphi(n) = \sum_{m=-N_w/N_h}^{N_w/N_h} \tilde{\psi}(n + mN_h)$
    $w_a(n) = \psi_a(n)/\sqrt{\varphi(n)}$
    $w_s(n) = \psi_s(n)/\sqrt{\varphi(n)}$
end for
```

Algorithm 2 MH-CMDCT

```
Parameters: $w_a(n)$, $N_w$, $N_h$
Input: The time-domain representation of signal $x(n)$
Output: The subband representation $X_{k,l}$
$\Delta = (N_w/2 + 1)/2$
for $l = 0, 1, \dots$ do
    $x_l(n) = x(n)w_a(n - lN_h)$
    $\tilde{x}_l(n) = x_l(n + lN_h)$
    for $k = 0, 1, \dots, N_w/2 - 1$ do
        $\vartheta(k) = \sum_{n=0}^{N_w-1} \tilde{x}_l(n) \exp\left(-j\pi n(2k+1)/N_w\right)$
        $X_{k,l} = 2\vartheta^*(k) \exp\left(-2j\pi(k + 0.5)\Delta/N_w\right)$
    end for
end for
```

Algorithm 3 IMH-CMDCT

```
Parameters: $w_s(n)$, $N_w$, and $N_h$
Input: The subband-domain representation $X_{k,l}$
Output: The time-domain representation of signal $x(n)$
$\Delta = (N_w/2 + 1)/2$
for $l = 0, 1, \dots$ do
    for $k = 0, 1, \dots, N_w/2 - 1$ do
      $\vartheta(k) = 0.5X_{k,l}^* \exp(2j\pi(k + 0.5)]\Delta/N_w)$
      $\vartheta(N_w - k - 1) = 0.5X_{k,l} \exp(-2j\pi(k + 0.5)\Delta/N_w)$
    end for
    for $n = 0, 1, \dots, N_w - 1$ do
      $\tilde{x}_l(n) = w_s(n) \sum_{k=0}^{N_w-1} \vartheta(k) \exp\left(j\pi n(2k+1)/N_w\right)$
    end for
    $x_l(n) = \tilde{x}_l(n - lN_h)$
end for
$x(n) = \sum_l x_l(n)$
```

2.2. System Model

Since we process the data in blocks and to make the arithmetic tractable, we assume that the expansion over time can be approximated by a shift in each block [11]. Let's assume that we use window size N_w with hop size N_h to calculate the subband representation. Then, in the absence of skew, the lth subband representation microphone signal is calculated from a block of samples given by $d\left((N_h l + n)/f_p\right)$ for $n = 0, 1, \dots N_w - 1$. However, in the presence of skew, the block that is used to calculate the subband representation would be $d\left((N_h l + n)/(f_p + \delta)\right)$ for $n = 0, 1, \dots N_w - 1$. It is shown in [11] that this block can be approximated by $d\left((N_h l + n)/f_p + N_h l\delta/f_p^2\right)$ for $n = 0, 1, \dots N_w - 1$. This motivates formulating the block skew in a state-space tracking framework where the state transition is determined by the clock skew. In this formulation, the state is the acoustic path between the loudspeaker and the microphone. Let $\boldsymbol{h}_{k,l}$ denote the channel at block l for subband k modeled as an FIR filter of length M. Then, the state equation with a vector valued function $\boldsymbol{\theta}(\cdot, \cdot)$ is given by

$$\boldsymbol{h}_{k,l+1} = \boldsymbol{\theta}(\boldsymbol{h}_{k,l}, \boldsymbol{u}_{k,l}) + \boldsymbol{\nu}_{k,l}, \tag{4}$$

where $\boldsymbol{u}_{k,l}$ is the control input (innovation) and $\boldsymbol{\nu}_{k,l}$ is the process noise with the covariance $\Gamma^2_{\Delta,k,l} := \mathcal{E}\{\boldsymbol{\nu}_{k,l}\boldsymbol{\nu}_{k,l}^H\}$ where $\mathcal{E}\{\cdot\}$ is the expectation and $\{\cdot\}^H$ denotes a Hermitian transpose. For the special case of linear Gauss-Markov model

$$\boldsymbol{h}_{k,l+1} = \boldsymbol{A}_k \boldsymbol{h}_{k,l} + \boldsymbol{\nu}_{k,l}, \tag{5}$$

where $\boldsymbol{A}_k$ represents a state transition matrix. The measurement equation is given by

$$d_{k,l} = \boldsymbol{h}_{k,l}^H \boldsymbol{x}_{k,l} + \varepsilon_{k,l}, \tag{6}$$

where $\varepsilon_{k,l}$ is the measurment noise for subband k and at block l with the covariance $\xi^2 := \mathcal{E}\{\varepsilon\varepsilon^*\}$ where $\{\cdot\}^*$ is a conjugate operator. We use an adaptive filter to the auxilary cost function

$$\mathcal{J}(\boldsymbol{a}_k) = \hat{\mathcal{E}}\left\{||\breve{\boldsymbol{h}}_{k,l+1} - \breve{\boldsymbol{h}}_{k,l}\boldsymbol{a}_k||^2_{\Gamma^2_\Delta}\right\}, \tag{7}$$

where $\|\cdot\|_{\mathbf{C}}$ is the ℓ_2-weighted norm with covariance matrix $\mathbf{C}$, $\boldsymbol{a}_k$ is a first column vector of $\boldsymbol{A}_k$ assuming the state transition is a convolution operation, and $\breve{\boldsymbol{h}}$ is a Toeplitz matrix which corresponds to the convolution with $\boldsymbol{h}$.

Remark 1. The Toeplitz constraint can be obtained by enforcing a diagonal structure on the estimated convolution matrix in the Fourier domain. At every adaptation step, $\hat{\boldsymbol{A}}_k$ is a full matrix from which only the diagonal elements are selected.

Remark 2. The oracle transition matrix $\boldsymbol{A}$ for all subbands when δ is known is given by

$$\boldsymbol{A} = \operatorname{diag}\left(\left[1, \phi^1, ..., \phi^{N_w-1}\right]\right), \tag{8}$$

where $\phi = \exp(j\frac{2\pi l\delta N_h}{N_w f_p})$, and $\operatorname{diag}(\boldsymbol{y})$ creates a diagonal matrix with the elements of vector $\boldsymbol{y}$. Once we calculate $\boldsymbol{A}_k$ then we use a Kalman filter which operates on (5) and (6).

Remark 3. Periodically, the accumulated delay becomes greater than a sample which is preserved as a sudden change in the channel. To solve this issue, we use a longer buffer for the far-end signal in low frequency subbands. Then, the peak value of estimated channel in those subbands is used to adjust the beginning far-end buffers for all the subbands.

3. Robust Kalman Filtering

Here, we show how to use a Kalman filtering approach to estimate both the acoustic path and the skew correction term in an alternating manner. A Kalman filter implements an optimal tracker in the least-squares sense by minimizing the following cost function with respect to $\boldsymbol{h}_{k,l}, \forall l$

$$\mathcal{J} = \hat{\mathcal{E}}\left\{||D_{k,l+1} - \boldsymbol{h}_{k,l+1}^H \boldsymbol{x}_{k,l+1}||^2_{\xi^2} + \lambda||\boldsymbol{h}_{k,l+1} - \boldsymbol{A}_k \boldsymbol{h}_{k,l}||^2_{\Gamma^2_\Delta}\right\},$$

where λ is the Lagrangian. It can be shown that minimization of this cost function can in-fact lead to a one-step point Kalman filter as follows, [12, 13]. We will omit the subband index dependency for notational clarity, we have

$$\boldsymbol{\mho}_l^2 = \left[\left(\boldsymbol{\Gamma}_{\Delta,l-1}^2 + \boldsymbol{A}_{l-1}\boldsymbol{\mho}_{l-1}^2\boldsymbol{A}_{l-1}^H\right)^{-1} + \xi_l^{-2}\boldsymbol{x}_l\boldsymbol{x}_l^H\right]^{-1} \tag{9}$$

$$\hat{\boldsymbol{h}}_l = \boldsymbol{A}_{l-1}\hat{\boldsymbol{h}}_{l-1} + \xi_l^{-2}\boldsymbol{\mho}_l^2\boldsymbol{x}_l \times \left[\left(\boldsymbol{A}_{l-1}\hat{\boldsymbol{h}}_{l-1}\right)^H \boldsymbol{x}_l - D_l\right],$$

where $\boldsymbol{\mho}_l^2$ denotes the inverse of the Hessian. Typically, direct implementation of the Kalman estimation suffers from instabilities and is sensitive to ill-conditioning. Better numerical properties are obtained by expressing the Kalman filter in the covariance form, [14]. The update-equation becomes

$$\begin{aligned}
\eta_{e,l}^2 &= \xi_l^2 + \boldsymbol{x}_l^H\boldsymbol{\mho}_{l-1}^2\boldsymbol{x}_l,\\
\boldsymbol{k}_l &= \eta_{e,l}^{-2}\boldsymbol{A}_l\boldsymbol{\mho}_{l-1}^2\boldsymbol{x}_l,\\
E_l &= D_l - \hat{\boldsymbol{h}}_l^H\boldsymbol{x}_l,\\
\hat{\boldsymbol{h}}_{l+1} &= \boldsymbol{A}_l\hat{\boldsymbol{h}}_l + \boldsymbol{k}_l E_l^*,\\
\boldsymbol{\mho}_l^2 &= \boldsymbol{A}_l\boldsymbol{\mho}_{l-1}^2\boldsymbol{A}_l^H + \Gamma_{\Delta,l}^2 - \eta_{e,l}^2\boldsymbol{k}_l\boldsymbol{k}_l^H.
\end{aligned} \tag{10}$$

Using the square root form, the update quantities are given by

$$\begin{bmatrix} \xi_l & \boldsymbol{x}_l^H\boldsymbol{\mho}_{l-1} & 0 \\ 0 & \boldsymbol{A}_l\boldsymbol{\mho}_{l-1} & \Gamma_{\Delta,l} \end{bmatrix} \boldsymbol{Q}_l = \begin{bmatrix} \eta_{e,l} & 0 & 0 \\ \bar{\boldsymbol{k}}_l & \boldsymbol{\mho}_l & 0 \end{bmatrix}, \tag{11}$$

where $\bar{\boldsymbol{k}}_l = \eta_{e,l}^2\boldsymbol{k}_l$, and $\boldsymbol{Q}_l$ is a unitary matrix designed to obtain a matrix with the triangular structure.

The Kalman filter is robust to Gaussian noise since it linear unbiased estimator in the presence of Gaussian noise. However, it is sensitive to non-Gaussian disturbances such as speech onsets and transient noises. One approach to increase robustness of the adaptive filters to super-Gaussian noises is to employ a Huber estimator, [15–17]. This method has been introduced to RLS and FRLS based filters. In the Kalman filter, the Gaussian bias term is already accounted for. In order to make a Kalman filter robust, one can use an estimation of the bias factor which follows a super-Gaussian distribution as introduced in [18]. We adopt this idea and augment the Kalman optimization criterion with the following auxiliary constraint.

$$\begin{aligned}
\mathcal{J}(\boldsymbol{\Upsilon}_k, r_k) =& \hat{\mathcal{E}}\left\{||D_{k,l+1} - \boldsymbol{x}_{k,l}^H\boldsymbol{h}_{k,l} - R_{k,l+1}||_{\xi^2}^2\right. \\
&\left. + \lambda||\boldsymbol{h}_{k,l+1} - \boldsymbol{A}\boldsymbol{h}_{k,l}||_{\Gamma_\Delta^2}^2 + \lambda_1\|R_{k,l+1}\|_p^p\right\},
\end{aligned} \tag{12}$$

where the distribution $P\left(R_{k,l}\right)$ is super-Gaussian.

$$R_{k,l} = \left(|\boldsymbol{e}_{k,l}| - \Xi\left(\boldsymbol{e}_{k,l}\right)\right)\exp(j\angle(\boldsymbol{e}_{k,l})), \tag{13}$$

where $\angle(\cdot)$ denotes the angle of a complex number, and

$$\Xi_{k,l}\left(\boldsymbol{e}_{k,l}\right) = \min\left\{\frac{|\boldsymbol{e}_{k,l}|}{\iota_{k,l-1}}, \kappa_0\right\}, \tag{14}$$

where κ_0 is a constant controlling the robustness of the algorithm [17], and $\iota_{k,l}$ is a scale factor which quantifies the error spread and can be estimated recursively with a forgetting factor α and a normalization constant β as:

$$\iota_{k,l} = \alpha_i\iota_{k,l-1} + \frac{(1-\alpha_i)}{\beta}\Xi\left(\boldsymbol{e}_{k,l}\right)\iota_{k,l}. \tag{15}$$

Algorithm 4 summarizes the steps for a robust Kalman filter which performs the skew correction. It assumes the state transition matrix $\boldsymbol{A}$ is available at every step. $\boldsymbol{A}$ is calculated in an alternating manner by estimating the channel using Eq. (7) and a multichannel Kalman filter implemented in the Fourier domain as described in Algorithm 5.

Algorithm 4 Robust Kalman Filtering in Square Root Form

```
Input: Loudspeaker signal $X_{k,l}$, microphone signal $D_{k,l}$, $\forall k, l$
Output: The subband estimate of the near-end $\boldsymbol{e}_l$
$\boldsymbol{h}_{k,-1} = \boldsymbol{0}$, $\boldsymbol{g}_{k,-1} = \boldsymbol{0}$, $\boldsymbol{\mho}_{k,0} = \boldsymbol{J}$, $\forall k$, $1 < \kappa$, $0 < \beta \leq 1$,
  $0<\alpha_i<1, 0<\alpha<1$    ▷ $\boldsymbol{J}$ is anti-diagonal identity matrix
for $l = 0, 1, \ldots$ do
    for $k \in \{0, 1, \ldots K-1\}$ do
        $\boldsymbol{x}_{k,l} = [X_{k,l-M+1}, \ldots, X_{k,l}]^T$    ▷ $\{\cdot\}^T$ denotes a
  transpose operation.
        $E_{k,l} = d_{k,l} - \boldsymbol{h}_{k,l-1}^H\boldsymbol{x}_{k,l}$
        $\xi_{k,l}^2 = \alpha\xi_{k,l-1}^2 + (1-\alpha)|E_{k,l}|^2$
        $\begin{bmatrix}\xi_{k,l} & \boldsymbol{x}_{k,l}^H\boldsymbol{\mho}_{k,l-1} & 0\\ 0 & \boldsymbol{A}_{k,l}\boldsymbol{\mho}_{k,l-1} & \Gamma_{\Delta,k,l}\end{bmatrix}\boldsymbol{Q}_{k,l} = \begin{bmatrix}\eta_{e,k,l} & 0 & 0\\ \bar{\boldsymbol{k}}_{k,l} & \boldsymbol{\mho}_{k,l} & 0\end{bmatrix}$
        $\boldsymbol{q}_{k,l} = \frac{1}{\eta_{k,l}^2+\epsilon}\bar{\boldsymbol{k}}_{k,l}$
        if $|E_{k,l}| < \kappa\iota_{k,l}$ then
            $\zeta_{k,l} = E_{k,l} - \left(|E_{k,l}| - \frac{|E_{k,l}|}{\iota_{k,l}+\epsilon}\right)\angle(E_{k,l})$
        else $\zeta_{k,l} = E_{k,l} - (|E_{k,l}| - \kappa_0)\angle(E_{k,l})$
        end if
        $\iota_{k,l} = \alpha_i\iota_{k,l-1} + (1-\alpha_i)|\zeta_{k,l}|/\beta$
        $\boldsymbol{h}_{k,l} = \boldsymbol{h}_{k,l-1} + \boldsymbol{q}_{k,l}\zeta_{k,l}^*$
    end for
    $\boldsymbol{e}_l = [E_{0,l}, \ldots, E_{K-1,l}]^T$
end for
```

4. Results

Through a set of experiments, we evaluate the performance of the proposed skew robust Kalman filtering approach (SRK). A Bartlett-Hann window is used for both the analysis and synthesis windows. The number of taps M is set to 10. To quantify the performance of the AEC duration the far-end single talk, we use echo return loss enhancement (ERLE) defined as $10\log_{10}\mathcal{E}\{d^2(n)\}/\mathcal{E}\{e^2(n)\}$ (dB). During double talk, we use the perceptual evaluation of speech quality (PESQ) score between the AEC output and the near-end ground truth [19]. In the next two experiments, we set N_w to 1280 and N_h to 160.

4.1. The Performance of AEC in the Absence of Skew

We first evaluate the baseline performance using the dataset from [20]. We use only the real recordings for which the loudspeakers nonlinearities are strong and the microphone signals have non-stationary background noise and near-end speech. We compare the result of SRK with the AECs based on partitioned block frequency domain adaptive filtering (PBFDAF) [21], partitioned block frequency domain Kalman filtering (PBFDKF) [22,23], and subband normalized least square (SB-NLMS) [24]. To control the AEC during the double talk, SB-NLMS uses an adaptive step size, [25]. PBFDAF uses the same mechanism with a background filter to safely copy the filter coefficients in the absence of double talk, [26]. To avoid divergence during double talk PBFDKF uses a variable step size which is controlled automatically by estimating the near-end signal, [22]. Table 1 shows SRK outperforms the other methods in both ERLE and PESQ score.

4.2. The Performance of AEC in the Presence of Skew

On the microphone signals in the dataset of Sect. 4.1, we apply a 2Hz clock skew which is equivalent to $\rho \approx 1.0001$ for $f_p =$ 16KHz. The result of using PBFDAF, PBFDKF, SB-NLMS,

Algorithm 5 Estimating the Skew Correction Term

```
Input: Loudspeaker signal $X_{k,l}$, microphone signal $D_{k,l}, \forall k, l$
Output: The estimate of the skew correction filter $\boldsymbol{A}_{k,l}$
$\underline{\boldsymbol{A}}_{k,-1} = \boldsymbol{0}$, $\underline{\boldsymbol{\mho}}_{k,0} = \boldsymbol{J}$, $\forall k$, $\underline{\gamma}_{k,-1,f} = 0, \forall k, f, 1 < \kappa, 0 < \alpha < 1, 0 < \lambda < 1$
for $l = 0, 1, \ldots$ do
    for $k \in \{0, 1, \ldots K-1\}$ do
        $\mathring{\boldsymbol{h}}_{k,l} = [\boldsymbol{h}_{k,l}^T \, 0 \ldots 0]^T$    ▷ Zero-padding
        $\underline{\boldsymbol{h}}_{k,l} = \mathcal{F}(\mathring{\boldsymbol{h}}_{k,l})$    ▷ $\mathcal{F}(\cdot)$ is an FFT operator.
        ▷ $\boldsymbol{h}_{k,l}$ is estimated without skew correction.
        for $f = 0, 1, \ldots, N_f - 1$ do ▷ $N_f$ is the FFT size.
            $\underline{\boldsymbol{e}}_{k,l,f} = \underline{\boldsymbol{h}}_{k,l,f} - \underline{\boldsymbol{a}}_{k,l-1,f}^H \underline{\boldsymbol{h}}_{k,l-1}$
            ▷ $\underline{\boldsymbol{a}}_{k,l,f}$ is the $f$-th column of matrix $\underline{\boldsymbol{A}}_{k,l}$.
            ▷ $\underline{\boldsymbol{h}}_{k,l,f}$ is the $f$-th element of $\underline{\boldsymbol{h}}_{k,l}$.
            $\underline{\gamma}_{k,l,f}^2 = \alpha \underline{\gamma}_{k,l-1,f}^2 + (1-\alpha) \left|\underline{\boldsymbol{e}}_{k,l}\right|^2$
            $\begin{bmatrix} \underline{\Upsilon}_{k,l} & \underline{\boldsymbol{h}}_{k,l-1}^H \underline{\boldsymbol{\mho}}_{k,l-1,f} \\ 0 & \lambda^{\frac{1}{2}} \underline{\boldsymbol{\mho}}_{k,l-1,f} \end{bmatrix} \boldsymbol{Q}_{k,l} = \begin{bmatrix} \eta_{e,k,l,f} & 0 \\ \underline{\bar{\boldsymbol{k}}}_{k,l,f} & \underline{\boldsymbol{\mho}}_{k,l,f} \end{bmatrix}$
            $\underline{\boldsymbol{q}}_{k,l,f} = \frac{1}{\eta_{k,l,f}^2 + \epsilon} \underline{\bar{\boldsymbol{k}}}_{k,l,f}$
            $\underline{\boldsymbol{a}}_{k,l,f} = \underline{\boldsymbol{a}}_{k,l-1,f} + \underline{\boldsymbol{q}}_{k,l,f} \underline{\boldsymbol{e}}^*(k, l, f)$
        end for
    end for
    $\underline{\boldsymbol{A}} = \text{diag}\{\underline{\boldsymbol{A}}\}_{k,l}$
    $\boldsymbol{A}_{k,l} = \mathcal{C}(\mathcal{F}^{-1}(\underline{\boldsymbol{A}})_{k,l})$
    ▷ $\mathcal{F}^{-1}(\cdot)$ is an inverse FFT operator and $\boldsymbol{P} = \mathcal{C}(\boldsymbol{p})$ converts a column vector to a covariance matrix and returns the convolution matrix $\boldsymbol{P}$ such that the product of $\boldsymbol{P}$ and a vector $\boldsymbol{q}$ is the convolution of $\boldsymbol{p}$ and $\boldsymbol{q}$
end for
```

Table 1: *Comparisons of Different AECs in the Absence of Skew*

	ERLE (dB)	PESQ
PBFDAF	7.69	1.48
PBFDKF	4.33	1.37
SB-NLMS	2.77	1.37
SRK	11.2	1.54

and SRK is shown in Table 2. SRK produces higher ERLE and PESQ score.

4.3. The Effect of Skew on AEC

In this experiment, we use a white noise as the far-end signal and by convolving it with a room impulse response we obtain the microphone signal. The clock skew of up to hundreds of parts per million (ppm) is common in consumer-grade system clocks [27]. We set $\rho \approx 1.001$ equivalent to 125 ppm. The AEC without skew compensation can only remove under 1dB of echo. However as shown in Fig. 1, our proposed method is able to suppress the echo significantly more.

4.4. The Effect of Hop Size on AEC

In this experiment, we evaluate the effect of hop size on the performance. We set N_w to 1280 and use different hop sizes of 160, 80, and 40. The result is shown in Fig. 2. As expected, the performance improves when reducing the hop size.

Table 2: *Comparisons of Different AECs in the Presence of Skew*

	ERLE (dB)	PESQ
PBFDAF	3.36	1.36
PBFDKF	3.96	1.31
SB-NLMS	1.56	1.29
SRK	6.57	1.53

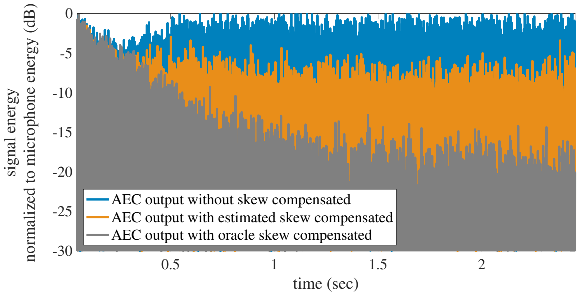

Figure 1: *The performance of SRK with and without skew compensation with the window size of* 1280 *and the hop size of* 160.

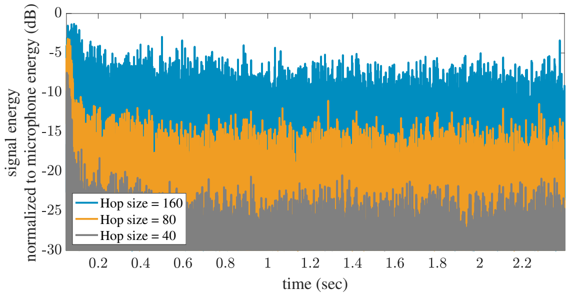

Figure 2: *The performance of our skew-compensated AEC with the window length of* 1280 *when we use different hop sizes.*

5. Conclusions

In this paper, we presented a novel state-space approach to acoustic echo cancellation which blindly accounts for the clock skew between playback and recording devices which are driven by independent clocks. The presented approach is implemented in a subband-domain based on a new variation of the complex modified discrete cosine transform which allows for arbitrary hop size and hence enhanced time resolution. Experimental results confirmed the effectiveness of the approach.

6. References

[1] J. Benesty, *Adaptive Estimation of Clock Skew and Different Types of Delay in the Internet Network*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 341–351.

[2] L. Wang and S. Doclo, "Correlation maximization-based sampling rate offset estimation for distributed microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 571–582, 2016.

[3] D. Cherkassky and S. Gannot, "Blind synchronization in wireless acoustic sensor networks," *IEEE/ACM Transactions on Au-*

dio, Speech, and Language Processing, vol. 25, no. 3, pp. 651–661, 2017.

[4] P. Thüne and G. Enzner, "Tracking theory of adaptive filters with input-output sampling rate offset," in *2019 27th European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[5] C. Avendano, "Acoustic echo suppression in the STFT domain," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*. IEEE, 2001, pp. 175–178.

[6] K. Helwani, H. Buchner, J. Benesty, and J. Chen, "A single-channel MVDR filter for acoustic echo suppression," *IEEE Signal Processing Letters*, vol. 20, no. 4, pp. 351–354, 2013.

[7] J.-M. Valin, S. Tenneti, K. Helwani, U. Isik, and A. Krishnaswamy, "Low-complexity, real-time joint neural echo control and speech enhancement based on PercepNet," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7133–7137.

[8] F.-L. Luo, *Mobile Multimedia Broadcasting Standards*. Springer, 2009.

[9] M. Mathew, V. Bhat, S. Thomas, and C. Yim, "Modified MP3 encoder using complex modified cosine transform," in *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, vol. 2, 2003, pp. II–709.

[10] S. F. Waldron, *An introduction to finite tight frames*. Springer, 2018.

[11] L. Wang and S. Doclo, "Correlation maximization-based sampling rate offset estimation for distributed microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 571–582, 2016.

[12] J. Humpherys and J. West, "Kalman filtering with Newton's method [lecture notes]," *IEEE Control Systems Magazine*, vol. 30, no. 6, pp. 101–106, 2010.

[13] H. Buchner, K. Helwani, and S. Godsill, "Blind signal processing for time-varying convolutive mixing systems based on sequence estimation on partly smooth manifolds," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7913–7917.

[14] S. Haykin, *Adaptive filter theory*, 4th ed. Upper Saddle River, NJ: Prentice Hall, 2002.

[15] P. J. Huber, *Robust statistics*. John Wiley & Sons, 2004, vol. 523.

[16] T. Gansler, S. L. Gay, M. M. Sondhi, and J. Benesty, "Double-talk robust fast converging algorithms for network echo cancellation," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 656–663, 2000.

[17] H. Buchner, J. Benesty, T. Gansler, and W. Kellermann, "Robust extended multidelay filter and double-talk detector for acoustic echo cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1633–1644, 2006.

[18] Y. Jin and B. D. Rao, "Algorithms for robust linear regression by exploiting the connection to sparse signal recovery," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 3830–3833.

[19] "ITU-T, P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001.

[20] R. Cutler, A. Saabas, T. Parnamaa, M. Purin, H. Gamper, S. Braun, K. Sorensen, and R. Aichner, "ICASSP 2022 acoustic echo cancellation challenge," in *ICASSP 2022*, 2022, the dataset uses other datasets licensed under CC0 license: https://research.google.com/audioset/index.html, https://freesound.org/, and https://zenodo.org/record/1227121#.XRKKxYhKiUk.

[21] J.-S. Soo and K. K. Pang, "Multidelay block frequency domain adaptive filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 2, pp. 373–376, 1990.

[22] F. Kuech, E. Mabande, and G. Enzner, "State-space architecture of the partitioned-block-based acoustic echo controller," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 1295–1299.

[23] T. Dietzen, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. van Waterschoot, "Partitioned block frequency domain kalman filter for multi-channel linear prediction based blind speech dereverberation," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2016, pp. 1–5.

[24] W. Kellermann, "Acoustic echo cancellation in subbands," *The Journal of the Acoustical Society of America*, vol. 87, no. S1, pp. S2–S2, 1990.

[25] J.-M. Valin, "On adjusting the learning rate in frequency domain echo cancellation with double-talk," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1030–1034, 2007.

[26] K. Ochiai, T. Araseki, and T. Ogihara, "Echo canceler with two echo path models," *IEEE Transactions on Communications*, vol. 25, no. 6, pp. 589–595, 1977.

[27] Y.-Y. Tai and M. F. Mansour, "Audio watermarking over the air with modulated self-correlation," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2452–2456.