applied sciences

MDPI

Article

Residual Echo Suppression Considering Harmonic Distortion and Temporal Correlation

Hyungchan Song and Jong Won Shin *

School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, 123 Cheomdan-gwagiro, Buk-gu, Gwangju 61005, Korea; shchan420@gist.ac.kr
* Correspondence: jwshin@gist.ac.kr

Received: 11 June 2020; Accepted: 29 July 2020; Published: 30 July 2020

Featured Application: Acoustic echo cancellation for speech communication, speech recognition and hearing aids.

Abstract: In acoustic echo cancellation, a certain level of residual echo resides in the output of the linear echo canceller because of the nonlinearity of the power amplifier, loudspeaker, and acoustic transfer function in addition to the estimation error of the linear echo canceller. The residual echo in the current frame is correlated not only to the linear echo estimates for the harmonically-related frequency bins in the current frame, but also with linear echo estimates, residual echo estimates, and microphone signals in adjacent frames. In this paper, we propose a residual echo suppression scheme considering harmonic distortion and temporal correlation in the short-time Fourier transform domain. To exploit residual echo estimates and microphone signals in past frames without the adverse effect of the near-end speech and noise, we adopt a double-talk detector which is tuned to have a low false rejection rate of double-talks. Experimental results show that the proposed method outperformed the conventional approach in terms of the echo return loss enhancement during single-talk periods and the perceptual evaluation of speech quality scores during double-talk periods.

Keywords: acoustic echo cancellation; residual echo suppression; harmonic distortion; temporal correlation

1. Introduction

Acoustic echo caused by acoustic coupling among microphones and loudspeakers is one of the most important issues in many applications of audio and speech signal processing such as speech communication [1], speech recognition [2], and hearing aids [3,4]. This allows the signal coming from the loudspeaker to be captured on the microphone, which degrades the quality of the speech communication or speech recognition rate. Acoustic echo cancellation (AEC) or acoustic echo suppression (AES) have been introduced to remove acoustic echoes [5–20]. Many AEC approaches employ linear adaptive filters in the time, frequency, or subband domain to predict and cancel out acoustic echoes based on far-end signals [5,9,13]. In the time domain, the most widely used method for linear AEC may be the normalized least mean square (NLMS) algorithm, possibly with step-size control [5] or a double-talk detector (DTD) [6,7], which provides a good balance between fast convergence and low misalignment. However, the length of the time domain adaptive filter should be long enough to accommodate possible impulse responses of the echo path, which sometimes requires huge computation. As alternative solutions, frequency domain and subband domain approaches have been proposed [8–13], which can reduce the computational complexity and increase the convergence speed simultaneously. In [9], AEC based on a frequency domain Kalman filter with a shadow filter approach employing an echo path change detector was proposed with reconvergence analysis. Ni et al. [13] proposed a combination of two subband adaptive filters with different step sizes without estimating system noise power, which showed fast convergence speed and small steady-state mean square error.

The linear filtering approach cannot completely remove the acoustic echo as the echo is not a linear function of the far-end signal. The nonlinearity arise mainly from the nonlinear response of the power amplifier and loudspeaker, as well as from the nonlinear acoustic transfer function and the misalignment of the linear echo canceller. In order to suppress nonlinear echo components, AEC based on a nonlinear adaptive filter has been proposed [14–17]. Volterra filter-based methods [14–16] were proposed to model the nonlinear relationship between far-end and acoustic echo signals with polynomial models. Unfortunately, these methods exhibited slow convergence rates [21]. Park et al. [17] also employed a polynomilal model in a Kalman filter-based framework when multiple microphone signals were available. AES algorithms analogous to speech enhancement techniques that estimates spectral gain functions such as Wiener filtering have also been proposed, and demonstrated an impressive performance with low computational complexity [18–20].

Another class of approach is placing a separate module after the linear echo canceller to clean up the residual echo left after the linear AEC or AES, which is called residual echo suppression (RES) [22–28]. Hoshuyama et al. [22] suggested a spectral subtraction scheme to remove the residual echo by assuming that the spectral magnitude of the residual echo is proportional to that of the linear echo estimate. Lee and Kim [23] proposed a statistical model based RES incorporating four hypotheses according to the existence of the near-end speech and residual echo. Schwarz et al. [24] proposed a RES that estimates residual echo from the far-end signal using an artificial neural network. In [25,26], RES based on deep neural networks have been proposed and have shown good performance, while requiring heavy computation. Another class of approaches is based on adaptive filters [27,28] that showed decent performance with a reasonable computational cost [21]. Bendersky et al. [28] proposed harmonic distortion RES (HDRES) in short-time Fourier transform (STFT) domain, which models the residual echo as a linear function of the linear echo estimates in the frequency bins that can make a harmonic distortion in the current frequency bin.

In this paper, we extend HDRES in [28] by modeling the residual echo in the current time-frequency bin as a function of the linear echo estimate, residual echo estimate, and microphone signals in adjacent frames as well as the linear echo estimates in the harmonically-related frequency bins in the current frame. A DTD is adopted to take account of residual echo estimates and microphone signals in the past frames without the adverse effect of the near-end speech and noise.

2. Problem Formulation

Let $x(t)$ denote the far-end signal and let $s(t)$ and $n(t)$ denote near-end speech and background noise with time index t. AEC output signal $e(t)$ and the microphone signal $y(t)$ are expressed as follows:

$$y(t) = d(t) + s(t) + n(t), \tag{1}$$

$$e(t) = \left(d(t) - \hat{d}(t)\right) + s(t) + n(t) = d_r(t) + s(t) + n(t), \tag{2}$$

in which $d(t)$ is the echo signal, $\hat{d}(t)$ is the linear echo estimate produced by an AEC filter, and $d_r(t)$ is the residual echo. The residual echo is always left in the output signal of the AEC, due to the nonlinearity arising from the power amps, loudspeakers, nonlinear echo path, and imperfect AEC. In the frequency domain, these equations become:

$$Y(m,f) = D(m,f) + S(m,f) + N(m,f), \tag{3}$$

$$\begin{aligned} E(m,f) &= \left(D(m,f) - \hat{D}(m,f)\right) + S(m,f) + N(m,f) \\ &= D_r(m,f) + S(m,f) + N(m,f), \end{aligned} \tag{4}$$

where $Y(m,f)$, $E(m,f)$, $D(m,f)$, $\hat{D}(m,f)$, $D_r(m,f)$, $S(m,f)$, and $N(m,f)$ are the STFT coefficients of $y(t)$, $e(t)$, $x(t)$, $d(t)$, $d_r(t)$, $s(t)$, and $n(t)$ in the frame m and frequency f, respectively. The goal of the RES is to estimate and remove the residual echo, $D_r(m,f)$, from the available signals, such as the far-end reference signal, $X(m,f)$, the linear echo estimate, $\hat{D}(m,f)$, and microphone signals, $Y(m,f)$, in the past and current frames in all frequency bins. A block diagram of an AEC system with a RES module is shown in Figure 1.

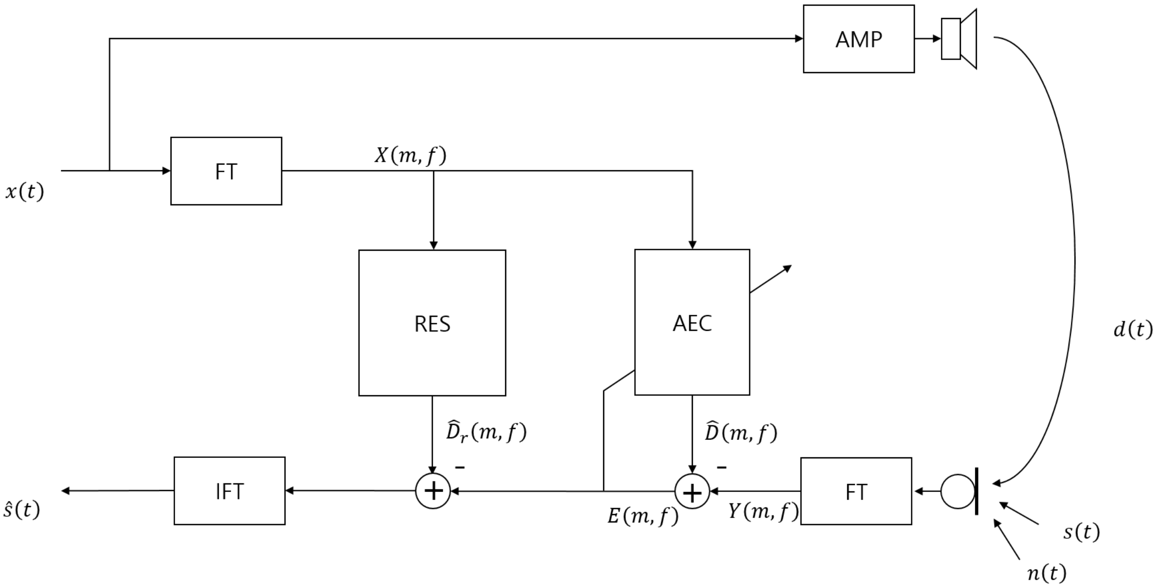

Figure 1. General block diagram of the acoustic echo cancellation system with residual echo suppression.

3. Harmonic Distortion Residual Echo Suppression

HDRES [28] estimates the magnitude of the residual echo as a linear function of the linear echo estimates in the frequency bins that can affect the current frequency bin with a harmonic distortion, i.e., the frequencies that are quotients of the current frequency and integers. With the $\delta(\cdot)$, which is 1 only for the frequency bins harmonically related to the current frequency and a few nearby bins, the estimate of the magnitude of the residual echo becomes:

$$|\hat{D}_r(m,f)| = \sum_{i=1}^{M}\sum_{j=1}^{H}\sum_{k=-K}^{K} \delta(i,j,k,f) W_H(i,j,k)|X'(m,i)|, \tag{5}$$

$$\delta(i,j,k,f) = \begin{cases} 1, & \text{if } i \times j + k = f \\ 0 & otherwise \end{cases}, \tag{6}$$

where M is the number of frequencies, H is the number of harmonics considered, K is the harmonic search window to accommodate nearby frequency bins to deal with the insufficient frequency resolution of the STFT, $W_H(i,j,k)$'s are weights of the linear combination, and $X'(m,f)$ is the linear echo estimate defined by [28]:

$$|X'(m,f)| = \sum_{t=0}^{S-1} L(t,f)|X(m-t,f)|, \tag{7}$$

where the normalized weighting factor $L(t,f)$ is:

$$L(t,f) = \frac{|W_L(t,f)|}{\sum_{j=0}^{S-1}|W_L(j,f)|}, \tag{8}$$

in which W_L is the weight of the frequency domain linear adaptive AEC filter [8,28] considering S frames. The function δ in (6) is constructed to deal with the harmonic distortion as the frequency contents falling in the i-th frequency bin affects the bins of which the frequencies are integer multiples of that frequency. The affected bins are centered on the bin $i \times j$ but may also include nearby $2K$ frequency bins, as the frequency resolution of the STFT is limited. The parameters of the residual echo suppression, $W_H(i,j,k)'$s, are estimated by a NLMS algorithm.

4. Residual Echo Suppression Considering Harmonic Distortion and Temporal Correlation

In this paper, we propose an extension of HDRES [28] which considers not only the harmonic distortion but also the temporal correlation of the relevant signals. The residual echo in the current time-frequency bin is modeled as a linear function of not only the linear echo estimates in the harmonically-related frequency bins in the current frame but also the linear echo estimate, residual echo estimate, and microphone signals in adjacent frames. Speech signal has strong temporal correlation that is beneficial to exploit in the vast areas of speech signal processing. In addition to the temporal correlation of the far-end speech signal or the filtered version of it that can help the estimation of the residual echo, the nonlinear part of the echo signal may have its own temporal correlation. Moreover, the microphone signals bears raw information in the captured signal that might be complementary to that in the estimated signals, which may also help the estimation of the residual echo in the current frame in the far-end single-talk regions.

To utilize all the relevant signals considering spectral and temporal correlation, two estimates for the spectral magnitude of the residual echo for far-end single-talk and double-talk periods are maintained. The residual echo for the far-end single-talk period in the frame m and frequency f is modeled as:

$$\begin{aligned}|\hat{D}_{ST}(m,f)| = &\sum_{i=1}^{M}\sum_{j=1}^{H}\sum_{k=-K}^{K} \delta(i,j,k,f)W_{Hs}(i,j,k)|X'(m,i)| + \sum_{p=1}^{T} W_{Ts1}(p,f)|X'(m-p,f)| \\ &+ \sum_{p=1}^{T} W_{Ts2}(p,f)|\hat{D}_r(m-p,f)| + \sum_{p=0}^{T} W_{Ts3}(p,f)|Y(m-p,f)|, \end{aligned} \tag{9}$$

where T is the number of considered previous frames, $W_{Hs}(i,j,k)$ is the weight for the linear echo estimates in the harmonically related frequencies and $W_{Ts1}(p,f)$, $W_{Ts2}(p,f)$, and $W_{Ts3}(p,f)$ are the weights for the linear echo estimates, residual echo estimates, and microphone signals in the previous frames, respectively. It is noted that the summation index for Y starts with 0 to allow the effect of the microphone signal in the current frame. On the other hand, the residual echo in the double-talk period is estimated without the microphone signal to avoid the adverse effect of the near-end signals:

$$\begin{aligned}|\hat{D}_{DT}(m,f)| = &\sum_{i=1}^{M}\sum_{j=1}^{H}\sum_{k=-K}^{K} \delta(i,j,k,f)W_{Hd}(i,j,k)|X'(m,i)| \\ &+ \sum_{p=1}^{T} W_{Td1}(p,f)|X'(m-p,f)| + \sum_{p=1}^{T} W_{Td2}(p,f)|\hat{D}_r(m-p,f)|, \end{aligned} \tag{10}$$

in which $W_{Hd}(i,j,k)$ is the weight for the linear echo estimates in the harmonically-related frequencies and $W_{Td1}(p,f)$ and $W_{Td2}(p,f)$ are the weights for the linear echo estimates, and residual echo estimates in the previous frames, respectively. With these two estimates, the estimate for the magnitude of the residual echo is determined depending on the result of the double-talk detection:

$$|\hat{D}_r(m,f)| = \begin{cases} |\hat{D}_{DT}(m,f)|, & \text{if double-talk is detected} \\ |\hat{D}_{ST}(m,f)|, & otherwise \end{cases}. \tag{11}$$

The weights in Equations (9) and (10) are updated during far-end single-talk periods using the NLMS algorithm that minimizes the mean square error between the AEC output signal and the residual echo estimate. The weights $W_{Hs}(i,j,k)$, $W_{T1s}(p,f)$, $W_{T2s}(p,f)$, and $W_{T3s}(p,f)$ used for the single-talk periods are adapted as follows:

$$W_{Hs}(i,j,k) \leftarrow W_{Hs}(i,j,k) + \frac{\mu}{\bar{P}_X(m,i)}|X'(m,i)|\xi_{ST}(m,i\times j+k), \tag{12}$$

$$W_{Ts1}(p,f) \leftarrow W_{Ts1}(p,f) + \frac{\mu}{\bar{P}_X(m-p,f)}|X'(m-p,f)|\xi_{ST}(m,f), \tag{13}$$

$$W_{Ts2}(p,f) \leftarrow W_{Ts2}(p,f) + \frac{\mu}{\bar{P}_D(m-p,f)}|\hat{D}_r(m-p,f)|\xi_{ST}(m,f), \tag{14}$$

$$W_{Ts3}(p,f) \leftarrow W_{Ts3}(p,f) + \frac{\mu}{\bar{P}_Y(m-p,f)}|Y(m-p,f)|\xi_{ST}(m,f), \tag{15}$$

$$\xi_{ST}(m,f) = |E(m,f)| - |\hat{D}_{ST}(m,f)|, \tag{16}$$

where $\xi_{ST}(m,f)$ is the error, μ is the step size, and $\bar{P}_X$, $\bar{P}_D$, and $\bar{P}_Y$ are smoothed powers given by:

$$\bar{P}_X(m,f) = (1-\rho)\bar{P}_X(m-1,f) + \rho|X'(m,f)|^2, \tag{17}$$

$$\bar{P}_D(m,f) = (1-\rho)\bar{P}_D(m-1,f) + \rho|\hat{D}_r(m,f)|^2, \tag{18}$$

$$\bar{P}_Y(m,f) = (1-\rho)\bar{P}_Y(m-1,f) + \rho|Y(m,f)|^2, \tag{19}$$

in which ρ is the smoothing parameter. As for the parameters used to estimate the residual noise during the double-talk period, $W_{Hd}(i,j,k)$, $W_{T1d}(p,f)$, and $W_{T2d}(p,f)$, are updated as follows:

$$W_{Hd}(i,j,k) \leftarrow W_{Hd}(i,j,k) + \frac{\mu}{\bar{P}_X(m,i)}|X'(m,i)|\xi_{DT}(m,i\times j+k), \tag{20}$$

$$W_{Td1}(p,f) \leftarrow W_{Td1}(p,f) + \frac{\mu}{\bar{P}_X(m-p,f)}|X'(m-p,f)|\xi_{DT}(m,f), \tag{21}$$

$$W_{Td2}(p,f) \leftarrow W_{Td2}(p,f) + \frac{\mu}{\bar{P}_D(m-p,f)}|\hat{D}_r(m-p,f)|\xi_{DT}(m,f). \tag{22}$$

$$\xi_{DT}(m,f) = |E(m,f)| - |\hat{D}_{DT}(m,f)|. \tag{23}$$

It is noted that the weights used to estimate the residual echo magnitude in the double-talk periods are updated with Equations (20)–(22) in the far-end single-talk period, not in the double-talk period. This is because the weights W_{Hd}, W_{Td1}, and W_{Td2} try to estimate the effect of the linear echo in the harmonically related frequencies, the linear echo in the previous frames, and the residual echo in the previous frames to the residual echo in the current frame, but the $|X'|$ and $|\hat{D}_r|$ in the double talk period contain a significant amount of near-end speech which disrupts the estimation of W_{Hd}, W_{Td1}, and W_{Td2}.

With the estimated magnitude of the residual echo in the Equation (11), $|\hat{D}_r(m,f)|$, the output of the linear echo canceller, $|E(m,f)|$, and the estimate of the magnitude of the noise spectrum obtained by the minimum statistics approach [29], $|\hat{N}(m,f)|$, the real-valued gain function of the residual echo suppressor, $G(m,f)$, is constructed as a spectral subtraction with a noise floor as in [27,28].

$$G(m,f) = \frac{\max(\bar{E}(m,f) - \beta\bar{D}_r(m,f), \bar{N}(m,f))}{\bar{E}(m,f)}, \tag{24}$$

where β is a parameter that controls the aggressiveness of the RES, $\max(\cdot,\cdot)$ function returns the largest value between two variables, and $\bar{D}_r$, $\bar{E}$, and $\bar{N}$ are smoothed versions of $|\hat{D}_r$, $|E|$, and $|\hat{N}|$ obtained by:

$$\bar{D}_r(m,f) = (1-\alpha)\bar{D}_r(m-1,f) + \alpha|\hat{D}_r(m,f)|, \tag{25}$$
$$\bar{E}(m,f) = (1-\alpha)\bar{E}(m-1,f) + \alpha|E(m,f)|, \tag{26}$$
$$\bar{N}(m,f) = (1-\alpha)\bar{N}(m-1,f) + \alpha|\hat{N}(m,f)|, \tag{27}$$

in which α is a smoothing factor. The final output of the RES is obtained by:

$$\hat{S}(m,f) = G(m,f)|E(m,f)|\exp\left(j\angle E(m,f)\right), \tag{28}$$

where $\exp(\cdot)$ is the exponential function, and the time domain signal $\hat{s}(t)$ is computed by an inverse short-time Fourier transform of $\hat{S}(m,f)$.

5. Experiments

To demonstrate the performance of the proposed RES, we compared the echo return loss enhancements (ERLEs) in the far-end single-talk periods and the perceptual evaluation of speech quality (PESQ) scores in double-talk periods for the linear AEC output without RES, that with the power filter-based RES (PFRES) [27], that with the HDRES [28], and that with the proposed RES. As for the linear AEC, we adopted the frequency-domain NLMS-based echo canceller proposed in [8]. The DTD module used in the experiments was the one proposed in [30]. The ERLE is defined by:

$$ERLE(t) = 10\log_{10}\left[\frac{E[y(t)^2]}{E[\hat{s}(t)^2]}\right]. \tag{29}$$

The parameter values for the proposed RES were $M = 257$, $S = 3$, $H = 5$, $K = 1$, $T = 1$, $\mu = 0.01$, $\rho = 0.9$, $\alpha = 0.7$, and $\beta = 2$. The parameters H, T, and K are related to the nonlinearity of the acoustic echo, so the values for those parameter would depend on the device configurations. The smoothing factors ρ and α were set to be rather high values, as the echo signal is highly nonstationary. The step size parameter μ was set to control the trade-off between the convergence rate and misalignment. We performed experiments for both simulated and real-recorded data with various noise types, echo-to-noise ratios (ENRs), near-end signal-to-noise ratios (SNRs), and near-end signal-to-echo ratios (SERs).

5.1. Experiments with Simulated Data

Firstly, the performances of the RES algorithms were evaluated with the simulated data. A total of 20 utterances spoken by 10 male and 10 female speakers were selected from the TIMIT database [31] as the far-end speech, while another 20 utterances from the same database were used as the near-end speech for the double-talk scenario. Background noises were Babble, White, and Factory noises from the NOISEX-92 database [32]. The sampling rates were 16 kHz. The ENR for the single-talk periods was set to be 10, 15, and 20 dB for each type of noise in addition to the noise-free condition. Therefore, the total number of data for the far-end single-talk scenario was $20 \times (3 \times 3 + 1) = 200$. As for the double-talk data, 20 pairs of utterances were used for near-end and far-end speech signals mixed at the SER of -5, 0, and 5 dB. The SNR was set to be 10, 15, and 20 dB for each noise type in addition to the clean noiseless condition, making the total number of data $20 \times 3 \times (3 \times 3 + 1) = 600$.

To simulate the nonlinearity that arise from power amplifiers and loudspeakers, we adopted a clipping function [33] and a memoryless sigmoidal function [34]. The hard clipping function [33] is defined as:

$$x_{hard}(t) = \begin{cases} -x_{max}, & x(t) < -x_{max} \\ x(t), & |x(t)| \leq x_{max} \\ x_{max}, & x(t) > x_{max} \end{cases}, \tag{30}$$

where x_{max} was set to be 80% of the maximum volume of $x(t)$. To model the nonlinear characteristics of the loudspeakers, the memoryless sigmoidal function [34] was used:

$$x_{NL}(t) = \gamma \cdot \left(\frac{2}{1 + \exp(-a \cdot b(t))} - 1 \right), \tag{31}$$

where

$$b(n) = 1.5 \times x(t) - 0.3 \times x(t)^2, \tag{32}$$

the sigmoid gain parameter γ is set to 2, and the sigmoid slope is set to be $a = 4$ if $b(t) > 0$ and $a = 0.5$ otherwise. The echo path after the nonlinearity was modeled with the image method [35] which simulates a 4 × 4 × 3 m small office with the reverberation time of $T_{60} = 200$ ms.

Tables 1 and 2 show the average ERLEs in the far-end single-talk periods and average PESQ scores in the double-talk periods for the simulated data. In all noise types and echo-to-noise ratios, the HDRES [28] improved the ERLE of the linear AEC while keeping the PESQ scores the same, and the proposed RES outperformed the HDRES in both the ERLE and PESQ scores. The PESQ scores for the PFRES was similar to those for HDRES, while the ERLE for the PFRES was in the middle of that for the HDRES and that for the proposed RES. On average, the ERLE was improved by 5.37 dB compared with the linear AEC output, by 0.74 dB compared with the PFRES, and by 4.43 dB compared with the HDRES, while the PESQ scores was improved by 0.064 compared with the linear AEC output, by 0.042 compared with the PFRES, and by 0.051 compared with the HDRES. We can conclude that exploiting the temporal correlation of the relevant signals in addition to the harmonic relation between frequency bins was effective to estimate the residual echo.

Table 1. Echo return loss enhancement (ERLEs) in various echo-to-noise ratio (ENR) conditions for the acoustic echo canceller (AEC) without residual echo suppression (RES), with the power filter-based RES (PFRES) [27], with the harmonic distortion RES (HDRES) [28], and with the proposed RES during the far-end single-talk periods for the simulated data. The numbers in the bold face indicate the best performances.

Noise Type	ENR	AEC [8,30]	AEC [8,30] +PFRES [27]	AEC [8,30] +HDRES [28]	AEC [8,30] +Proposed RES
clean		8.59	13.14	9.58	**14.41**
White	20 dB	7.94	12.56	8.96	**13.53**
	15 dB	7.34	12.03	8.29	**12.52**
	10 dB	6.25	10.85	7.09	**11.04**
Babble	20 dB	7.87	12.71	8.89	**13.53**
	15 dB	7.30	12.07	8.26	**12.67**
	10 dB	6.21	10.86	7.07	**11.32**
Factory	20 dB	7.90	12.36	8.91	**13.55**
	15 dB	7.36	11.91	8.32	**12.78**
	10 dB	6.28	10.92	7.13	**11.44**
Average		7.31	11.94	8.25	**12.68**

Table 2. Average perceptual evaluation of speech quality (PESQ) scores in various signal-to-noise ratio (SNR) and signal-to-echo ratio (SER) conditions for the AEC without RES, with the PFRES [27], with the HDRES [28], and with the proposed RES during the double-talk periods for the simulated data. The numbers in the bold face indicate the best performances.

Noise Type	SNR	SER	AEC [8,30]	AEC [8,30] +PFRES [27]	AEC [8,30] +HDRES [28]	AEC [8,30] +Proposed RES
clean		5dB	2.902	2.911	2.918	**2.991**
		0 dB	2.665	2.710	2.697	**2.771**
		−5 dB	2.468	2.511	2.473	**2.540**
White	20 dB	5 dB	2.660	2.658	2.673	**2.734**
		0 dB	2.459	2.490	2.491	**2.553**
		−5 dB	2.287	2.309	2.291	**2.345**
	15 dB	5 dB	2.498	2.485	2.510	**2.567**
		0 dB	2.351	2.362	2.362	**2.421**
		−5 dB	2.179	2.200	2.183	**2.227**
	10 dB	5 dB	2.279	2.270	2.290	**2.340**
		0 dB	2.179	2.181	2.189	**2.241**
		−5 dB	2.039	2.045	2.041	**2.079**
Babble	20 dB	5 dB	2.501	2.512	2.513	**2.568**
		0 dB	2.322	2.368	2.352	**2.407**
		−5 dB	2.170	2.208	2.174	**2.217**
	15 dB	5 dB	2.322	2.336	2.332	**2.381**
		0 dB	2.180	2.227	2.208	**2.255**
		−5 dB	2.055	2.085	2.057	**2.092**
	10 dB	5 dB	2.089	2.108	2.096	**2.138**
		0 dB	1.991	2.043	2.016	**2.053**
		−5 dB	1.904	1.926	1.904	**1.929**
Factory	20dB	5dB	2.584	2.589	2.598	**2.658**
		0 dB	2.391	2.432	2.422	**2.479**
		−5 dB	2.228	2.254	2.231	**2.280**
	15 dB	5 dB	2.422	2.430	2.434	**2.489**
		0 dB	2.259	2.306	2.289	**2.342**
		−5 dB	2.118	2.150	2.122	**2.166**
	10 dB	5 dB	2.206	2.212	2.216	**2.266**
		0 dB	2.090	2.126	2.117	**2.161**
		−5 dB	1.985	2.007	1.987	**2.022**
Average			2.293	2.315	2.306	**2.357**

5.2. Experiments with Real-Recorded Data

Since the nonlinear echo may not be simulated well enough by the clipping function and memoryless sigmoidal function, we additionally performed experiments with real recordings. We used 28 far-end speech signals and a near-end speech signal recorded with a commercial mobile phone, Samsung Galaxy S8, in hand-held hands-free mode [36]. The raw microphone signals and the far-end signal are obtained by an internal development program in Samsung. Each data have a length of about 65 seconds with a sampling rate of 16 kHz. A total of 3 types of noises including Pub, Road, and Call center noises from the European Telecommunications Standards Institute (ETSI) EG 202 396-1 background noise database [37], were replayed from the loudspeakers in the recording room to simulate background noises. The total number of real-recorded data for the far-end single-talk and double-talk periods considering the same ENR, SNR, and SER conditions with simulated data were $28 \times (3 \times 3 + 1) = 280$ and $28 \times 3 \times (3 \times 3 + 1) = 840$.

Table 3 shows the average ERLE with real-recorded data for the far-end single-talk period. The average ERLE for the proposed RES was 5.23 dB higher than that of the linear AEC output, 1.27 dB higher than that of the PFRES, and 1.53 dB higher than that of the HDRES. Utilizing the correlation with the microphone signal and the previous frames of the linear echo estimate, microphone signal, and residual echo estimates were shown to be effective in the real recordings with a commercial smartphone, too. Figure 2 demonstrates an example of the ERLEs over time for the AEC system without RES, with the PFRES, with the HDRES, and with the proposed RES along with the microphone signal waveform. In Table 4, the PESQ scores for real recordings are shown. Again, we can confirm that the proposed RES could achieve better speech quality in the double-talk periods in various background noise and echo conditions for the real-recorded data.

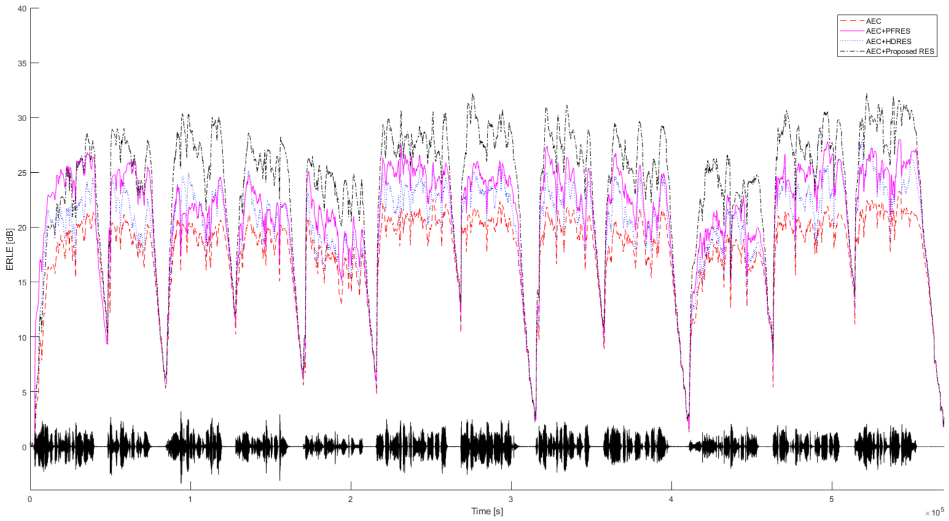

Figure 2. The ERLEs of the AEC system without RES, with PFRES, with HDRES, and with the proposed RES evolving with time and the microphone signal for one real-recorded signal with far-end single-talk without background noise.

Table 3. ERLEs in various ENR conditions for the AEC, with the PFRES [27], with the HDRES [28], and with the proposed RES during the far-end single-talk periods for the real-recorded data. The numbers in the bold face indicate the best performances.

Noise Type	ENR	AEC [8,30]	AEC [8,30] +PFRES [27]	AEC [8,30] +HDRES [28]	AEC [8,30] +Proposed RES
clean		13.44	20.58	20.38	**22.39**
Pub	20 dB	13.27	17.29	17.15	**18.39**
	15 dB	11.24	14.73	14.94	**15.90**
	10 dB	8.51	11.61	12.04	**13.00**
Road	20 dB	13.44	17.63	16.96	**18.93**
	15 dB	11.56	15.36	14.69	**16.65**
	10 dB	9.02	12.51	11.77	**13.86**
Callcenter	20 dB	13.47	17.45	17.06	**18.53**
	15 dB	11.69	15.13	14.93	**16.21**
	10 dB	9.21	12.19	12.01	**13.30**
Average		11.49	15.45	15.19	**16.72**

Table 4. Average PESQ scores in various SNR and SER conditions for the AEC without RES, with the PFRES [27], with the HDRES [28], and with the proposed RES during the double-talk periods for the real-recorded data. The numbers in the bold face indicate the best performances.

Noise Type	SNR	SER	AEC [8,30]	AEC [8,30] +PFRES [27]	AEC [8,30] +HDRES [28]	AEC [8,30] +Proposed Method
clean		5 dB	3.178	3.206	3.243	**3.275**
		0 dB	3.013	3.057	3.087	**3.135**
		−5 dB	2.815	2.874	2.879	**2.962**
Pub	20 dB	5 dB	3.005	3.024	3.033	**3.065**
		0 dB	2.910	2.944	2.953	**3.000**
		−5 dB	2.747	2.799	2.792	**2.871**
	15 dB	5 dB	2.767	2.788	2.779	**2.807**
		0 dB	2.719	2.753	2.740	**2.780**
		−5 dB	2.603	2.653	2.623	**2.694**
	10 dB	5 dB	2.684	2.707	2.694	**2.720**
		0 dB	2.642	2.675	2.657	**2.696**
		−5 dB	2.543	2.593	2.554	**2.624**
Road	20 dB	5 dB	3.097	3.123	3.144	**3.174**
		0 dB	2.969	3.010	3.029	**3.074**
		−5 dB	2.781	2.838	2.838	**2.913**
	15 dB	5 dB	2.904	2.935	2.940	**2.969**
		0 dB	2.824	2.867	2.871	**2.913**
		−5 dB	2.681	2.736	2.724	**2.797**
	10 dB	5 dB	2.828	2.861	2.864	**2.892**
		0 dB	2.758	2.802	2.804	**2.846**
		−5 dB	2.637	2.693	2.675	**2.748**
Callcenter	20 dB	5 dB	3.056	3.078	3.092	**3.127**
		0 dB	2.944	2.983	2.994	**3.043**
		−5 dB	2.769	2.824	2.817	**2.897**
	15 dB	5 dB	2.836	2.858	2.858	**2.889**
		0 dB	2.776	2.810	2.806	**2.851**
		−5 dB	2.641	2.693	2.674	**2.747**
	10 dB	5 dB	2.750	2.774	2.769	**2.800**
		0 dB	2.702	2.736	2.728	**2.772**
		−5 dB	2.589	2.638	2.618	**2.687**
Average			2.806	2.844	2.843	**2.892**

5.3. Computational Complexity of the Proposed RES Algorithm

We investigated the computational complexity of the proposed RES algorithm. The proposed RES algorithm in our experiment requires $\{2(2K+1)H + 5T + S + 46\}M$ real-valued multiplications per frame, while HDRES algorithm [28] requires $\{(2K+1)H + S + 22\}M$ real-valued multiplications and the PFRES algorithm [27] requires $(6p^2 + 13p + 3)M$ real-valued multiplications, where p denotes the order of the polynomial model of the nonlinear echo path. For the parameter values we used in the experiments (e.g., $M = 257$, $T = 1$, $S = 3$, $H = 5$, $K = 1$, and $p = 3$), the computational complexity of the proposed method is approximately twice of that of the HDRES [28] and 10% less than that of the PFRES [27]. We can see that the proposed method has reasonable computational complexity considering the performance improvement shown in Tables 1–4.

6. Conclusions

In this paper, we proposed a method for residual echo suppression considering harmonic distortion and temporal correlation. The proposed method estimates residual echo taking account of not only the linear echo estimates in the harmonically-related bins but also the linear echo estimates, residual echo estimates, and the microphone signals in adjacent frames. To utilize the microphone signal without the adverse effect of the near-end signals, the DTD module is utilized. Experimental results showed that the proposed method improved the ERLE of the HDRES in the far-end single talk period by 4.43 dB for the simulated data and 1.53 dB for the real-recorded data, and the PESQ scores of the HDRES in the double-talk periods by 0.051 for both simulated data and 0.049 for the real-recorded data, which may justify the increase of the computational complexity.

Author Contributions: Conceptualization, H.S. and J.W.S.; methodology, J.W.S.; software, H.S.; validation, J.W.S.; formal analysis, J.W.S.; investigation, H.S.; resources, J.W.S.; data curation, H.S.; writing—original draft preparation, H.S.; writing—review and editing, J.W.S.; visualization, H.S.; supervision, J.W.S.; project administration, J.W.S.; funding acquisition, J.W.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Research Foundation of Korea grant number NRF-2019R1A2C2089324 and Samsung System LSI (SLSI-201801GE002S).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gay, S.L.; Benesty, J. *Acoustic Signal Processing for Telecommunication*; Springer Science & Business Media: Berlin, Germany, 2012; Volume 551.
2. Hong, J. Stereophonic Acoustic Echo Suppression for Speech Interfaces for Intelligent TV Applications. *IEEE Trans. Consum. Electron.* **2018**, *64*, 153–161. [CrossRef]
3. Spriet, A.; Proudler, I.; Moonen, M.; Wouters, J. Adaptive feedback cancellation in hearing aids with linear prediction of the desired signal. *IEEE Trans. Signal Process.* **2005**, *53*, 3749–3763. [CrossRef]
4. Spriet, A.; Rombouts, G.; Moonen, M.; Wouters, J. Adaptive feedback cancellation in hearing aids. *J. Frankl. Inst.* **2006**, *343*, 545–573. [CrossRef]
5. Paleologu, C.; Ciochină, S.; Benesty, J.; Grant, S.L. An overview on optimized NLMS algorithms for acoustic echo cancellation. *EURASIP J. Adv. Signal Process.* **2015**, *2015*, 97. [CrossRef]
6. Jung, H.K.; Kim, N.S.; Kim, T. A new double-talk detector using echo path estimation. *Speech Commun.* **2005**, *45*, 41–48. [CrossRef]
7. Gänsler, T.; Benesty, J. The fast normalized cross-correlation double-talk detector. *Signal Process.* **2006**, *86*, 1124–1139. [CrossRef]
8. Malvar, H. A modulated complex lapped transform and its applications to audio processing. In Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP99 (Cat. No. 99CH36258), Phoenix, AZ, USA, 15–19 March 1999; IEEE: Piscataway, NJ, USA, 1999; Volume 3, pp. 1421–1424.
9. Yang, F.; Enzner, G.; Yang, J. Frequency-domain adaptive Kalman filter with fast recovery of abrupt echo-path changes. *IEEE Signal Process. Lett.* **2017**, *24*, 1778–1782. [CrossRef]
10. Shi, K.; Ma, X. A frequency domain step-size control method for LMS algorithms. *IEEE Signal Process. Lett.* **2009**, *17*, 125–128.
11. Yang, F.; Cao, Y.; Wu, M.; Albu, F.; Yang, J. Frequency-Domain Filtered-x LMS Algorithms for Active Noise Control: A Review and New Insights. *Appl. Sci.* **2018**, *8*, 2313. [CrossRef]
12. Ni, J.; Li, F. A variable step-size matrix normalized subband adaptive filter. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *18*, 1290–1299.
13. Ni, J.; Li, F. Adaptive combination of subband adaptive filters for acoustic echo cancellation. *IEEE Trans. Consum. Electron.* **2010**, *56*, 1549–1555. [CrossRef]
14. Stenger, A.; Trautmann, L.; Rabenstein, R. Nonlinear acoustic echo cancellation with 2nd order adaptive Volterra filters. In Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP99 (Cat. No. 99CH36258), Phoenix, AZ, USA, 15–19 March 1999; Volume 2, pp. 877–880.

15. Guérin, A.; Faucon, G.; Le Bouquin-Jeannès, R. Nonlinear acoustic echo cancellation based on Volterra filters. *IEEE Trans. Speech Audio Process.* **2003**, *11*, 672–683. [CrossRef]
16. Azpicueta-Ruiz, L.A.; Zeller, M.; Figueiras-Vidal, A.R.; Arenas-García, J.; Kellermann, W. Adaptive combination of Volterra kernels and its application to nonlinear acoustic echo cancellation. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *19*, 97–110. [CrossRef]
17. Park, J.; Chang, J.H. State-Space Microphone Array Nonlinear Acoustic Echo Cancellation Using Multi-Microphone Near-End Speech Covariance. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1520–1534. [CrossRef]
18. Avendano, C. 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), New Platz, NY, USA, 24–24 October 2001; pp. 175–178.
19. Faller, C.; Chen, J. Suppressing acoustic echo in a spectral envelope space. *IEEE Trans. Speech Audio Process.* **2005**, *13*, 1048–1062. [CrossRef]
20. Park, Y.S.; Chang, J.H. Frequency domain acoustic echo suppression based on soft decision. *IEEE Signal Process. Lett.* **2008**, *16*, 53–56. [CrossRef]
21. Panda, B.; Kar, A.; Chandra, M. Non-linear adaptive echo supression algorithms: A technical survey. In Proceedings of the 2014 International Conference on Communication and Signal Processing, Melmaruvathur, India, 3–5 April 2014; pp. 76–80.
22. Hoshuyama, O.; Sugiyama, A. An acoustic echo suppressor based on a frequency-domain model of highly nonlinear residual echo. In Proceedings of the 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Toulouse, France, 14–19 May 2006; Volume 5, p. 269-272.
23. Lee, S.Y.; Kim, N.S. A statistical model-based residual echo suppression. *IEEE Signal Process. Lett.* **2007**, *14*, 758–761. [CrossRef]
24. Schwarz, A.; Hofmann, C.; Kellermann, W. Spectral feature-based nonlinear residual echo suppression. In Proceedings of the Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 20–23 October 2013; pp. 1–4.
25. Lee, C.M.; Shin, J.W.; Kim, N.S. DNN-based residual echo suppression. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015.
26. Carbajal, G.; Serizel, R.; Vincent, E.; Humbert, E. Multiple-input neural network-based residual echo suppression. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 231–235.
27. Kuech, F.; Kellermann, W. Nonlinear residual echo suppression using a power filter model of the acoustic echo path. In Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP '07, Honolulu, HI, USA, 15–20 April 2007; Volume 1, pp. I–73.
28. Bendersky, D.A.; Stokes, J.W.; Malvar, H.S. Nonlinear residual acoustic echo suppression for high levels of harmonic distortion. In Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 31 March–4 April 2008; pp. 261–264.
29. Martin, R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech Audio Process.* **2001**, *9*, 504–512. [CrossRef]
30. Park, Y.S.; Chang, J.H. Double-talk detection based on soft decision for acoustic echo suppression. *Signal Process.* **2010**, *90*, 1737–1741. [CrossRef]
31. Lamel, L.F.; Kassel, R.H.; Seneff, S. Speech database development: Design and analysis of the acoustic-phonetic corpus. In Proceedings of ESCA Tutorial and Research Workshop on Speech Input/Output Assessment and Speech Databases, Noordwijkerhout, The Netherlands, 20–23 September 1989; pp. 161–170.
32. Varga, A.; Steeneken, H.J. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Commun.* **1993**, *12*, 247–251. [CrossRef]
33. Malik, S.; Enzner, G. State-space frequency-domain adaptive filtering for nonlinear acoustic echo cancellation. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *20*, 2065–2079. [CrossRef]
34. Comminiello, D.; Scarpiniti, M.; Azpicueta-Ruiz, L.A.; Arenas-Garcia, J.; Uncini, A. Functional link adaptive filters for nonlinear acoustic echo cancellation. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 1502–1512. [CrossRef]
35. Allen, J.B.; Berkley, D.A. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am.* **1979**, *65*, 943–950. [CrossRef]

36. 3GPP. *3GPP TS 26.132, Technical Specification Group Services and System Aspects*; Speech and Video Telephony Terminal Acoustic Test Specification; European Telecommunications Standards Institute: Sophia Antipolis, France, 2020.
37. ETSI. *ETSI EG 202 396-1 Speech Processing, Transmission and Quality Aspects (STQ)*; Speech Quality Performance in the Presence of Background Noise; Part 1: Background Noise Simulation Technique and Background Noise Database; European Telecommunications Standards Institute: Sophia Antipolis, France, 2008.