Integrated Near-End Acoustic Echo and Noise Reduction Systems

S. M. Kuo$^1$, D. W. Sun$^1$, and W. S. Gan$^2$

$^1$Dept. of Electrical Engineering  
Northern Illinois University  
DeKalb, IL 60115

$^2$School of Elec. & Electronics Engineering  
Nanyang Technological University  
Singapore (639798)

Abstract – This paper presents the development of integrated near-end acoustic echo and noise reduction algorithm. The modified frequency-sampling filter (FSF) provides an effective filterbank for splitting signal into equally spaced frequency channels. The center clipper and attenuator are employed at each frequency bin for attenuating near-end acoustic echo and noise. The adaptive clipping threshold and dynamic attenuator value are updated based on the optimized voice-activity detector (VAD). Computer simulations using speech and measured room impulse response show that the near-end acoustic echo and noise are suppressed successfully.

I. INTRODUCTION

The wide spread use of cellular phones has significantly increased the use of communication devices in high noise environments. Intense background noise, however, often corrupts speech, degrading the performance of communication systems. Existing digital signal processing techniques such as speech coding, automatic speech recognition, speaker identification, channel transmission, and echo cancellation are usually developed under noise-free assumption. These techniques could be employed in noisy environments if a front-end noise reduction algorithm can sufficiently reduce an additive noise. The noise reduction is becoming increasingly important with the development of hands-free and voice-activated cellular phones.

Traditional acoustic echo cancellation deals with the acoustic echo generated by the far-end talker, which is broadcasted in a room using speakerphones or hands-free phones. In general, a room or a vehicle compartment is a reverberated chamber. When the near-end talker inside the room talks, far-end listeners located at a remote location will receive not only the desired direct path energy, but also many delayed replicas with varying amplitudes, the so-called reverberation. This undesired effect makes the near-end speech sound hollow. Depending on the microphone location, the energy of these near-end acoustic echoes may be large enough to degrade the intelligibility of speech. Near-end acoustic echo reduction is critical for teleconferencing, multimedia, public addressing, mobile telephony, audio system correction, recorded signal quality improvement, speech recognition, and many other voice communications applications. In this paper, we focus on acoustic echoes produced by the near-end talker. This is different from the traditional acoustic echo cancellation, which focuses on the acoustic echo produced by the far-end talker through the loudspeaker in the room.

Signal processing techniques [1] can be applied to reduce acoustic noise and reverberation picked up by the microphone. Acoustic noise reduction and de-reverberation can be classified as speech enhancement problems. However, previous research works in this field treats these two problems separately. To solve these problems effectively and provide a quality that is sufficient for telecommunications, combined reduction of these disturbances is required.

The de-reverberation method developed by Allen [2] was based on voice production model. In this method, the reverberated speech was first analyzed to estimate the linear predictive coding parameters of clean speech, and then the de-reverberated speech was reconstructed. In reference [3], the adaptive algorithms in conjunction with other techniques were used in improving the intelligibility of corrupted speech in the reverberated environment. Based on the similarity of smoothed spectral magnitudes of measured room impulse responses, Cole and Moody proposed the spectral-subtraction-type method for the enhancement of reverberated speech [4]. The idea of using spectral subtraction is adapted and improved by using center clipper in this paper for reducing reverberation.

There are four speech enhancement techniques; each has its own set of assumptions, advantages, and limitations [5]. The first class of technique is based on the short-term spectrum. These techniques suppress noise by subtracting the noise spectrum estimated during non-speech activity. The second class of technique is based on speech modeling using iterative methods. These systems estimate speech parameters based on autoregressive or autoregressive-moving-average models, followed by re-synthesis of the noise-free speech using the non-causal Wiener filtering. The third class of system is based on adaptive noise canceling (ANC) using a dual-channel system with the least-mean-square algorithm. The last area of speech enhancement is based on the periodicity of voiced speech. These methods employ fundamental frequency tracking using either single-channel ANC or adaptive comb filtering of the harmonic
noise. In this research, we modified spectral subtraction algorithm [6] by estimating the short-term spectral magnitude of the noise and noisy speech spectra through the FSF instead of using fast Fourier transform.

The objective of this paper is to develop and integrate digital signal processing techniques for reducing acoustic noise and near-end acoustic echoes. The algorithm consists of four elements: a frequency-sampling filter, a voice activity detector, a dynamic attenuator, and an adaptive center clipper. Attenuator reduces the noise and center clipper removes the reverberated echoes. These two operations are performed at every frequency channel formed by the frequency-sampling filter [7].

II. FREQUENCY-SAMPLING FILTER

The FSF is based on sampling a desired amplitude spectrum and obtaining filter coefficients. In this approach, the desired frequency response \( H(\omega) \) is first uniformly sampled at \( L \) equally spaced points \( \omega_k = \frac{2 \pi k}{L}, \ k = 0, 1, \ldots, L-1 \). The frequency sampling technique is particularly useful when several bandpass functions are desired simultaneously. Another unique attraction of the frequency sampling method is that it allows recursive implementation of filters, leading to computationally efficient algorithms.

The transfer function of FSF can be expressed as [1]

\[
H(z) = \frac{2}{L} \left[ 1 - r z^{-1} \right] \sum_{k=1}^{L} H_k \frac{1 - r \cos(2\pi k / L) z^{-1}}{1 - 2r \cos(2\pi k / L) z^{-1} + r^2 z^{-2}},
\]

where \( r \) is the pole radius that is slightly less than one, and \( H_k \) is the gain at frequency channel \( k \). As illustrated in Fig. 1, the FSF can be realized by cascading a comb filter with a bank of second-order resonators.

In the FSF structure shown in Fig. 1, each resonator effectively acts as a narrowband filter and passes only the frequencies centered at and close to the resonant frequency. By choosing the coefficient \( H_k \) according to given specifications, a time-domain filter with an arbitrary magnitude response can be obtained.

III. VOICE ACTIVITY DETECTOR

Near-end acoustic echo and noise suppression algorithms based on the FSF have a unique way to distinguish the speech and non-speech by evaluating the power in some specific channels. It is well known that voiced speech has considerable energy at the first formant (250-800 Hz) and a majority of noises do not have strong resonant frequencies in this region. Based on this principle, an optimized VAD using an adaptive noise floor threshold scheme is developed in this paper.

The decision whether a given signal is considered as speech or non-speech is made based on the power in the first formant region. A dynamic noise floor estimate \( \hat{P}_f(n) \) used for estimating the background noise level is updated using a very long window (4 seconds) during the speech periods, but a median window (32 ms) is used for the non-speech periods. An adaptive threshold \( T(n) \) for the speech/non-speech decision is based on the time-varying noise floor estimate. A decision that the speech signal is present is made if the power estimate \( \hat{P}_f(n) \) exceeds the adaptive threshold \( T(n) \).

IV. ACOUSTIC NOISE ATTENUATION

The FSF structures discussed in Section II can be used to replace FFT in the speech enhancement technique based on spectral subtraction. In this case the input signal is split into \( K \) channels by the FSF instead of fast Fourier transform. Speech enhancement algorithm is applied to determine the dynamic attenuator value \( H_k(n) \) based on the power of noise and noisy speech, and these processed channel signals are processed by the center clipper for removing reverberation, and then recombined to form the overall output signal.

The block diagram of the FSF-based speech enhancement system is illustrated in Fig. 2. The noise and noisy signal levels are individually estimated in each band based on VAD output. The noise reduction (NR) algorithm is identical at each channel and is summarized as follows:

1. Estimate the power of noisy speech signal \( \hat{P}_{f,d}(n) \) for speech segments at each channel as follows:
\[
\hat{P}_{f,k}(n) = (1-\alpha)\hat{P}_{f,k}(n-1) + \alpha \hat{D}^2(n),
\]
where \(0 < \alpha < 1\).

2. Estimate the power of noise for noise segments at each channel as follows:
\[
\hat{P}_{f,k}(n) = (1-\beta)\hat{P}_{f,k}(n-1) + \alpha \hat{D}^2(n),
\]
where \(0 < \beta < 1\).

3. Compute dynamic attenuator value \(H_k(n)\) for each channel \(k\) using the principle of spectral subtraction algorithm expressed as follows:
\[
H_k(n) = 1 - \frac{\hat{P}_{f,k}(n)}{\hat{P}_{f,k}(n)}. \tag{4}
\]

4. Attenuate noise at each channel as follows:
\[
g_k(n) = \hat{f}_k(n)H_k(n). \tag{5}
\]
This channel output is then used by the center clipper for further reducing near-end acoustic echoes.

V. NEAR-END ACOUSTIC ECHO REDUCTION

For near-end acoustic echo reduction, the center clipper after dynamic attenuator shown in Fig. 2 is an effective technique to suppress undesired echo components. In general, there are two different kinds of center clippers. Two popular clipping algorithms are Center clipper A, which chops the signal components of the near-zero region (central region) and reduces the signal amplitude of non-central region with the desired amplitude. Center clipper B leaves the non-central region unchanged. The most popular clipper B can be expressed as
\[
y_k(n) = \begin{cases} 
0, & |g_k(n)| \leq C_k(n) \\
g_k(n), & |g_k(n)| > C_k(n) 
\end{cases}
\]
where \(C_k(n)\) is the adaptive clipping level. This center clipper completely eliminates signals below the clipping level, but leaves instantaneous signal values greater than the clipping level unaffected. Thus large signals go through unchanged but small signals are eliminated. Since small signals are consistent with reverberated elements, the clipping achieves the function of reducing near-end acoustic echoes.

The performance of near-end acoustic echo reduction can be further evaluated in frequency domain. Figure 5 shows the comparison of the spectrograms of the original speech (a), the reverberated speech (b), and de-reverberated speech (c). Figure 5(b) shows the near-end acoustic echoes blurred the gaps between the original speeches shown in Fig. 5(a). Figure 5(c) showed the center clipper effectively cleans up the blurred regions. Comparing the playback of original, reverberated, and processed signals shows that reverberated signal components are located at the tail portions of voice samples, \(C_k(n)\) can be further optimized by updating \(\beta(n)\) according to VAD result. The parameter \(\beta(n)\) takes larger value when a silence sample is detected and takes smaller value when a voice sample is presented.

VI. SIMULATION RESULTS

For simulation purpose, a room impulse response was measured in a conference room using the maximum-length sequence technique. The clean speech is then convoluted with the measured room impulse response to simulate the signal picks up by a microphone in that conference room. In this way, we have the original speech to evaluate the success of near-end acoustic echo reduction. The reverberated signal consists of two major components: the desired speech directly from the near-end talker and the undesired near-end echoes reflected from the walls and floor. As shown in Fig. 3, the shaded line (yellow) represents the reverberated speech; and the dark line (blue) represents the original speech.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig3.png}
\caption{Original speech (blue dark line) and the reverberated speech (yellow line).}
\end{figure}

The technique given in Section V was used to reduce the near-end acoustic echoes embedded in the desired speech. The FSF with \(L = 256\) and \(r = 0.995\) was used for simulations. The result is shown in Fig. 4, where the reverberated speech is shown as shaded line (yellow) and the dark line (blue) shows the near-end acoustic echoes were reduced. Comparing the de-reverberated signal (blue line in Fig. 4) with the original signal (blue line in Fig. 3), we showed that the near-end acoustic echo was significantly reduced, and the nonlinear distortion caused by the center clipper was minimized.

The performance of near-end acoustic echo reduction can be further evaluated in frequency domain. Figure 5 shows the comparison of the spectrograms of the original speech (a), the reverberated speech (b), and de-reverberated speech (c). Figure 5(b) shows the near-end acoustic echoes blurred the gaps between the original speeches shown in Fig. 5(a). Figure 5(c) showed the center clipper effectively cleans up the blurred regions. Comparing the playback of original, reverberated, and processed signals shows that reverberated
speech is suppressed significantly without noticeable distortion.

Fig. 4 De-reverberated speech (blue line) and reverberated speech (yellow line).

Fig. 5 Spectrograms of the (a) original, (b) reverberated, and (c) de-reverberated speech using center clipper.

For evaluating the performance of acoustic noise reduction introduced in Section IV, the noisy speech was recorded in a machinery room. As shown in the top graph of Fig. 6, the desired speech is embedded in broadband noise. The acoustic noise reduction algorithm was used to reduce the undesired noise. The enhanced speech is shown at the bottom plot of Fig. 6. Comparing the top and bottom waveforms, especially during the non-speech periods, we show that the algorithm able to reduce broadband noise significantly. Subjective evaluation by listening the original noisy speech and the enhanced speech shows the distortion of speech quality is very low.

Fig. 6 Performance of acoustic noise reduction. The original noisy speech is shown on the top graph, and the enhanced signal is displayed at the bottom graph.

REFERENCES: