MULTILAYER ADAPTATION BASED COMPLEX ECHO CANCELLATION AND VOICE ENHANCEMENT

Jun Yang (Senior Member, IEEE)

Amazon Lab126, 1100 Enterprise Way, Sunnyvale, CA 94089, USA
Email: junyang@amazon.com

ABSTRACT

The paper proposes an efficient signal processing system mainly consisting of an adaptation-based nonlinear echo cancellation (NLEC) layer and a joint perceptual subband residual echo suppression (SBRES) layer and noise reduction (SBNR) layer. The theoretical analyses, subjective and objective test results show that the proposed signal processing system can offer a significant improvement for automatic speech recognition and full-duplex voice communication performance in emerging artificial intelligence speakers. The proposed SBRES and NLEC layers can reduce various types of echoes including linear, nonlinear, and time-variant echo. Correspondingly, the proposed SBNR layer can effectively reduce not only noises but also echoes that have the similar statistical characteristics to noises. Non-uniform auditory perceptual critical bands are employed so as to better reflect cochlea mechanisms. The SBRES and SBNR layers are jointly accomplished in frequency domain, which results in a significant reduction of MIPS consumption from real time implementation point of view.

Index Terms **—** Nonlinear echo cancellation system, noise reduction, adaptive filters, automatic speech recognition, full-duplex voice communication

1. INTRODUCTION

For the purpose of improving automatic speech recognition (ASR) performance and full-duplex voice communication (FDVC) performance, acoustical echo cancellation (AEC) and noise reduction systems are playing a more important role in many emerging hands-free applications where noises and echoes are becoming more and more complex. A current AEC scheme usually employs an adaptive linear filter in either time domain, or frequency domain, or subband domain to model or approximate the real acoustic echo path between loudspeaker and microphone, and subtracts the estimated echo from the microphone signal.

However, there is actually always a residual echo after the above linear adaptive subtraction. This is due to the following reasons: (1). adaptive linear filter can neither be perfectly accurate nor exactly model the transfer function of the echo path, (2). the length of adaptive linear filter is not often sufficient. (3). there might be non-linearity in the echo path which is impossible for adaptive linear filter to model. Therefore, a nonlinear processor technique is necessary to further reduce the residual echo. On the other hand, the traditional nonlinear processors (such as, center clipper, noise-gate, spectral subtraction approaches, or other spectral enhancement techniques) will distort the near-end voice [1 - 7]. More importantly, an unnatural sounding residual echo can be produced if these existing nonlinear processor schemes are directly employed. This is mainly because of the following factors: (1). the user movement results in the echo path change, (2). the loudspeaker volume changes result in the time-varying echo, especially when the echo path changes faster than the convergence rate of adaptive linear filter, (3). adaptive linear filter could incorrectly "adjust" itself, which results in a reduction of near-end voice during the period when the near-end user is talking.

In practical applications, what makes the processing more challenging is the mixed situation where various echoes and noises simultaneously present.

Obviously, techniques that can efficiently suppress these various types of complex echoes and noise are highly desirable. To achieve this goal, this paper proposes a multilayer processing system, which mainly includes a joint perceptual SBRES layer and SBNR layer as well as an adaptation-based NLEC layer. The given theoretical analyses, subjective and objective test results show that the proposed system can offer a significant improvement for ASR and FDVC performance in emerging artificial intelligence speakers.

The rest of this paper is organized into the following four sections. Section 2 mainly presents the proposed algorithms of joint SBRES layer and SBNR layer. Section 3 presents the proposed adaptation-based NLEC layer. By using various test results, Section 4 mainly shows that the artificial intelligence speakers implemented with the proposed system can have significant improvements in terms of ASR and echo-return-loss-enhancement (ERLE) performance with good voice quality in real-time FDVC. Section 5 will make some conclusions and further discussions.

2. THE PROPOSED JOINT SBRES AND SBNR ALGORITHMS

The processing architecture of the proposed multilayer system is shown in Figure 1 with single-channel being example but without losing generality. In other words, this system is easily extended to the multiple-channel cases. In playback/receive path (i.e., Rx), the AVC standing for automatic volume control and limiter algorithms are proposed and implemented in [8], the EQ is an equalizer to compensate for loudspeaker frequency response.

In transmit path (i.e., Tx) of Figure 1, the block "Microphone" could be a single microphone or a microphone array for FDVC and ASR applications, respectively. The "Adaptive Linear AEC" is an existing echo preprocessor by using adaptive linear filter (ALF). The proposed processing of joint SBRES and SBNR layers is shown in Figure 2. The details of the proposed adaptation-based NLEC layer will be described in Section 3. AGC is an existing automatic gain control for voice communication.

To obtain "AEC reference", a sampling-rate-converter (SRC) is used. The "Rx HPF" and "Tx HPF" are of the same characteristics to remove frequencies lower than 80 Hz.

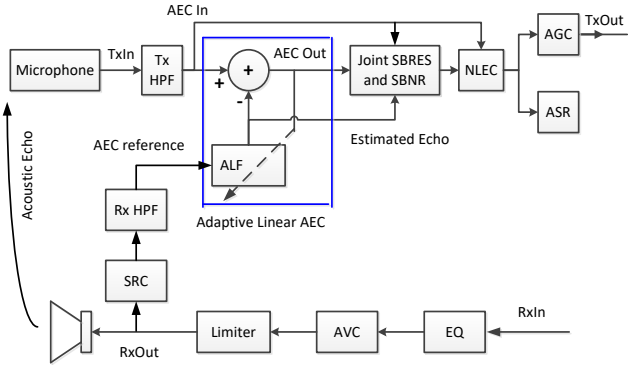

Figure 1 The Proposed Multilayer Processing System

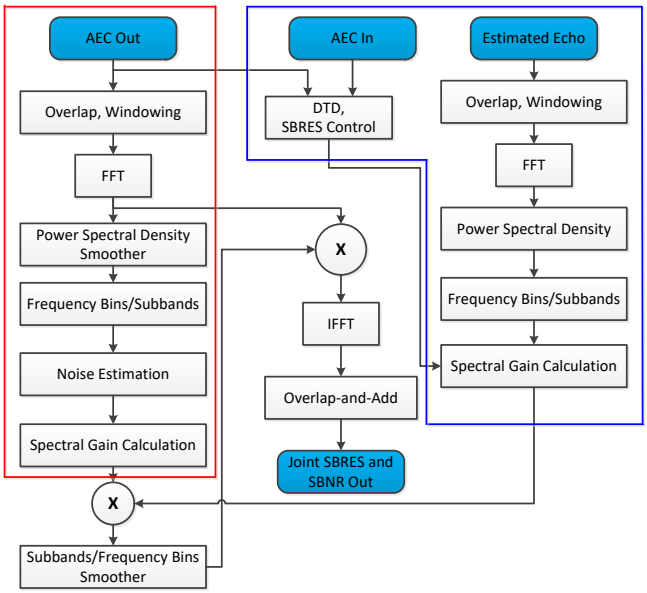

Figure 2 The Proposed Scheme for Joint SBRES and SBNR

In Figure 2, the blocks included in the red box belong to SBNR layer, the blocks included in blue big box belong to SBRES layer. The "Overlap" could be 50% between consecutive frames which is described as follows.

$$x(m,n) = x(m-1, L+n) \qquad 0 \le n < L \qquad (1)$$

where m is the current frame, n is the sample index, L is the number of audio samples in a frame, e.g., $L = 128$ samples for the configuration of 8 ms frame length and 16 kHz sampling rate. The $x(m, n)$ for $L \le n < 2L$ are the current audio samples of "AEC Out".

In Figure 2, the "Windowing" can be implemented by Hamming or Hanning function shown in Eq. (2), or the raised cosine function. Hanning function is as follows.

$$w(n) = 0.5(1.0 - \cos(2\pi n / N)) \qquad 0 \le n \le N-1 \qquad (2)$$

where N is the window length in number of audio samples. The $N = 2L$ for 50% overlap. The FFT is implemented by

$$X(k) = \frac{2}{N}\sum_{n=0}^{N-1} x(m,n)w(n)e^{-j2\pi nk/N} \qquad 0 \le k < N \qquad (3)$$

The "Power Spectra Density" (PSD) is $|X(k)|^2$ for $0 \le k \le L$, where $k=0$ denotes for DC component, k = L denotes for Nyquist component.

The two "smoother" blocks have the same processing and are implemented by a finite-impulse-response (FIR) low-pass filter. They are designed to smooth raw PSD and the obtained spectral bin gain over frequency.

The block "Frequency Bins/Subbands" converts from $(L+1)$ bins to either 30 or 15 non-uniform bands on the basis of the auditory critical bands.

Instead of relying on voice activity detection or speech presence probability, the proposed "Noise Estimation" algorithm stores the band PSD of the selected frame into a noise history window and estimates noise PSD from this PSD window by searching the minimum band PSD for each frequency band over a moving time window.

Without employing traditional parametric spectral subtraction, the proposed "Spectral Gain Calculation" has improved the Ephraim and Malah suppression rule in a global optimal way for both echo and noise in each frequency band. This processing could also output the optional voice activity detection information if needed by other processing parts.

The "DTD and SBRES Control" is the proposed double-talk-detector (DTD). Two DTD schemes are proposed. Both schemes can be performed in either subband or full-band domains. As an example, DTD_1 and DTD_2 are performed in subband and full-band domain, respectively. The DTD_1 is based on the cross-correlation between the "AEC In" signal $y(n)$ and "Estimated Echo" signal $z(n)$. The cross-correlation coefficients of each frequency band j in the m-th frame is defined as follows.

$$C_{yz}(m,j) = \frac{P_{yz}^2(m,\ j)}{P_y(m,\ j)P_z(m,\ j)} \qquad (4)$$

where $P_{yz}(m,j)$, $P_y(m,j)$ and $P_z(m,j)$ are cross-power and power estimations, respectively, and are defined as follows.

$$P_{yz}(m,j) = (1-\alpha)P_{yz}(m-1,j) + \alpha y(m,j)z(m,j) \qquad (5)$$

$$P_y(m,j) = (1-\alpha)P_y(m-1,j) + \alpha y^2(m,j) \quad (6)$$

$$P_z(m,j) = (1-\alpha)P_z(m-1,j) + \alpha z^2(m,j) \quad (7)$$

where α is a constant between 0 and 1. If the cross-correlation coefficient $C_{yz}(m, j)$ is less than a first threshold, then $DTD_1(m, j)$ = true, otherwise, $DTD_1(m, j)$ = false.

The proposed DTD_2 is based on ERLE measure of "Adaptive Linear AEC". The ERLE is calculated as follows.

$$ERLE = 10 * \log_{10}\left(\frac{E\{|y(n)|^2\}}{E\{|x(n)|^2\}}\right) \quad (8)$$

where E{} is the expectation operator. What $y(n)$ and $x(n)$ denote are audio samples of "AEC In" and "AEC Out", respectively. If ERLE is less than a second threshold, the DTD_2 = true, otherwise, DTD_2 = false.

Combining DTD_1 and DTD_2, a final DTD is determined. When the final DTD is determined as true, SBRES is dynamically disabled. Otherwise, SBRES is automatically enabled.

A "smoother" technique is applied to the spectral bin gain after combining the obtained spectral band gains of noise with that of echo and converting the final spectral band gain into spectral bin gain. Furthermore, the output complex spectrum is obtained after performing frequency domain filtering by applying the obtained optimal spectral bin gain to the input complex spectrum. An IFFT processing is performed to map the result from frequency domain to time domain. Then, the "Overlap-and-Add" approach is used to reconstruct a frame of samples; therefore, the noise and residual echo can be greatly suppressed and the processed output is also of high voice quality. It can be seen from the above that the proposed SBRES can reduce not only linear echo but also nonlinear echo. Also, the proposed SBNR can reduce not only noise but also stationary echo.

3. THE PROPOSED ADAPTATION-BASED NLEC ALGORITHM

The proposed adaptation-based NLEC layer is shown as in Figure 3, where the "Delay" should be the algorithm latency of the "Joint SBRES and SBNR" block so as to time-align the "AEC In" signal and "Joint SBRES and SBNR Out" signal. The proposed NLEC algorithm takes "AEC In" signal as reference which includes all types of echo nonlinearities.

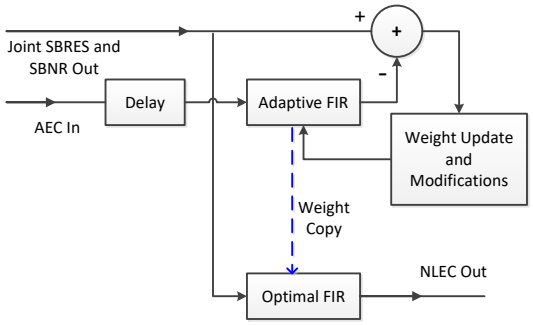

Figure 3 The Proposed Adaptation-Based NLEC Algorithm

The normalized least mean square (NLMS) adaptation scheme is used to update the weights $\boldsymbol{h}(n)$ of "Adaptive FIR" filter and is implemented as follows.

$$\boldsymbol{h}(n+1) = \boldsymbol{h}(n) + \frac{\mu \boldsymbol{e}(n)\boldsymbol{v}(n)}{\boldsymbol{v}^T(n)\boldsymbol{v}(n)} \quad (9)$$

where $\boldsymbol{e}(n)$ is the output of the "Adder", i.e., error signal. What $\boldsymbol{v}(n)$ denotes is the delayed "AEC In" signal, i.e., the reference signal with $\boldsymbol{v}^T(n)$ denoting its transpose. The step size of the adaptation is denoted by μ, whose value is between 0 and 1.

Instead of switching between freezing or unfreezing the adaptation in the conventional adaptive filtering algorithm, what this paper proposes is to dynamically adjust the filter weights after the adaptation. As shown in the "Weight Modification" of Figure 3, all the related weights are adjusted according to the three situations, i.e., double talk, near-end talk only, and far-end talk only.

More importantly, the proposed NLEC algorithm introduces a globally optimal FIR filter in addition to an adaptive FIR filter so as to maximize the performance of NLEC as further discussed in next sections. The "Weight Copy" contains a set of various measures that attempt to ascertain the convergence state of the two FIR filters.

4. EVALUATIONS

In this section, the evaluation results and test analyses of the proposed system are presented in terms of noise reduction performance, echo suppression performance, ASR performance, and FDVC performance.

4.1. Noise reduction performance

Figure 4 shows the input waveform (top) of noisy speech captured in vacuum noise environment and the output waveform (bottom) processed by the proposed SBNR layer. Obviously, the proposed SBNR layer reduces noise about 19.3 dB. Figure 5 shows the corresponding spectrograms.

Figure 4 Waveforms of "before" (top) and "after" (bottom) SBNR Processing

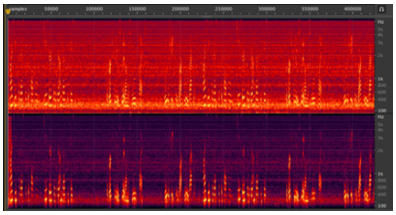

Figure 5 Spectrograms of Figure 4

4.2. Echo suppression performance

Figure 6 shows the waveform of SBRES=off (top) and the waveform of SBRES=on (bottom). It can be seen that the proposed SBRES layer reduces echo about 11.64 dB. Figure 7 shows the waveform of NLEC=off (top) and the waveform of NLEC=on (bottom), which shows that echo has been reduced by the proposed NLEC layer about 35 dB.

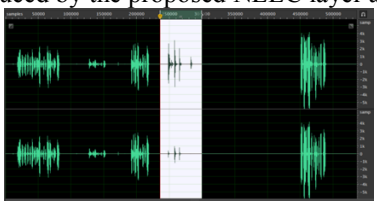

Figure 6 Waveforms of "SBRES=off" (top) and "SBRES=on" (bottom)

Figure 7 Waveforms of "NLEC=off" (top) and "NLEC=on" (bottom)

4.3. Full-duplex voice communication performance

Figure 8 shows the waveform of (SBRES, SBNR, NLEC) = off (top) and the waveform of (SBRES, SBNR, NLEC) = on (bottom). It can be seen from this result that the proposed (SBRES, SBNR, NLEC) reduces echo about 40 dB.

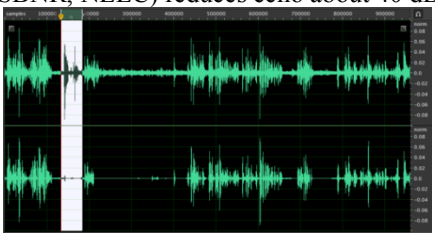

Figure 8 Waveforms of (SBRES, SBNR, NLEC) = off (top) and (SBRES, SBNR, NLEC) = on (bottom)

4.4. ASR performance

The ASR test results of the proposed SBRES, SBNR, and NLEC layers are obtained by using a third-party ASR engine. Figure 9 shows the relative word-error-rate (WER) improvement of SBNR layer for male (top plot) and female (bottom plot) voice, where averaging over 12 types of noises is performed. There are 6,000 utterances for each types of noises. Figure 10 shows the WER reductions of SBRES and SBNR layers with 19*1486 = 28,234 words. Figure 11 shows the relative WER improvements of NLEC layer. There are 10,000 wake-words for each playback volume.

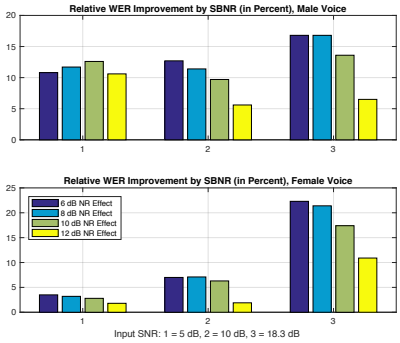

Figure 9 Relative WER Improvements of SBNR Layer

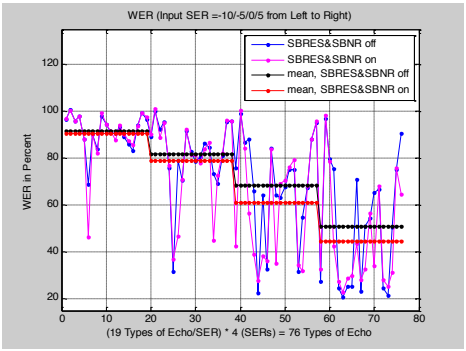

Figure 10 WER (lower is better) of SBRES and SBNR Layers

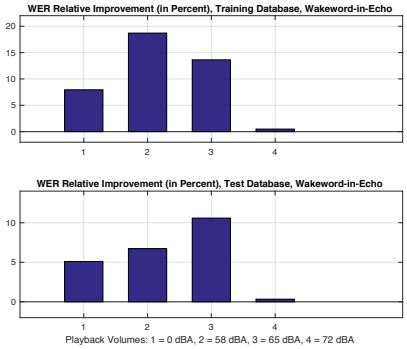

Figure 11 Relative WER Improvements of NLEC Layer

5. CONCLUSONS

By addressing various types of echoes and noises, the above theoretical analyses, subjective and objective test results have shown that the proposed signal processing system can offer a significant improvement for ASR and FDVC performance in emerging artificial intelligence speakers.

In addition, the MIPS requirement incurred by the proposed system is also small from real time implementation point of view. All of these mean that the proposed system can serve as a very efficient voice enhancement tool for many emerging audio/voice related applications and devices where echoes and noises are becoming complex and mixed.

6. REFERENCES

[1] Maria Luis Valero, Ilkay Yildiz, Edwin Mabande, and Emanuel A. P. Habets, “Coherence-Aware Stereophonic Residual Echo Estimation,” 2017 Hands-free Speech Communications and Microphone Arrays (HSCMA 2017), San Francisco, California, USA, pp. 176 - 180, March 1 - 3, 2017

[2] Jie Xia, Yi Zhou, and Ruitang Mao, “An Improved Cross-correlation Spectral Subtraction Post-processing Algorithm for Noise and Echo Canceller,” 2016 IEEE International Conference on Digital Signal Processing (DSP), Beijing, China, pp. 114 - 118, Oct. 16 - 18, 2016.

[3] Ingo Schalk-Schupp, Friedrich Faubel, Markus Buck, Andreas Wendemuth, “Approximation of a Nonlinear Distortion Function for Combined Linear and Nonlinear Residual Echo Suppression,” 2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC 2016), Xi’an, China, Sept. 13 - 16, 2016.

[4] Jason Wung, "A System Approach to Multi-Channel Acoustic Echo Cancellation and Residual Echo Suppression for Robust Hands-free Teleconferencing," Ph.D. Dissertation, School of Electrical and Computer Engineering, Georgia Institute of Technology, May 2015.

[5] Jason Wung, Ted S. Wada, Biing-Hwang Juang, Bowon Lee, Ton Kalker, and Ronald W. Schafer, “A System Approach to Residual Echo Suppression in Robust Hands-free Teleconferencing,” ICASSP 2011, Prague, Czech Republic, pp. 445 - 448, May 22 - 27, 2011.

[6] Urmila Shrawankar and Vilas Thakare, “Noise Estimation and Noise Removal Techniques for Speech Recognition in Adverse Environment,” Intelligent Information Processing V, 340, Springer, pp.336-342, 2010, IFIP Advances in Information and Communication Technology, 978-3-642-16326-5.

[7] Joon-Hyuk Chang, Hyoung-Gon Kim, and Sangki Kang, “Residual Echo Reduction Based on MMSE Estimator in Acoustic Echo Canceller,” 2007 IEICE Electronics Express, Vol. 4, No. 24, pp. 762-767, December 25, 2007.

[8] Jun Yang, Philip Hilmes, Brian Adair, and David W. Krueger, "Deep Learning Based Automatic Volume Control and Limiter System," ICASSP 2017, New Orleans, USA, pp. 2177 - 2181, March 5 - 9, 2017.