MULTILAYER ADAPTATION BASED COMPLEX ECHO CANCELLATION AND VOICE ENHANCEMENT

Jun Yang, Ph. D.

Amazon Lab126, Sunnyvale, CA 94089, USA

ICASSP 2018
April 15 - 20, 2018

Overview

- Voice over Internet Protocol (VoIP) and Automatic Speech Recognition (ASR) enabled artificial intelligence (AI) speaker is the killer application.
- Nonlinear Echo Cancellation (NLEC), Residual Echo Suppression (RES), and Noise Reduction (NR) are key parts in VoIP and ASR based AI speakers.
- Existing NLEC, RES, and NR are designed independently and hence result in poor ASR and full-duplex voice communication performance.
- By using perceptual subband (SB) architecture, a novel joint SBRES and SBNR algorithm is proposed.
- By using advanced adaptive filtering approach, a novel NLEC algorithm is proposed.
- The proposed SBRES, SBNR, and NLEC are integrated with a linear acoustic echo cancellation (AEC).
- Linear, nonlinear, and time-variant echo can be significantly reduced by SBRES, SBNR, and NLEC layers.
- Noise and stationary echo can be significantly reduced by SBNR layer.
- Optimal VoIP and ASR performance can be achieved.

The Proposed Joint SBRES and SBNR Alg. and NLEC Alg.

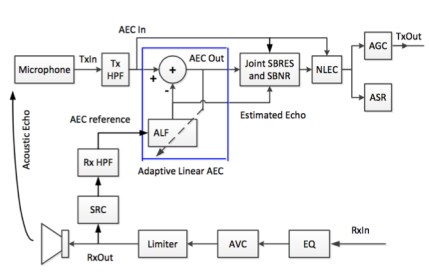

Fig. 1 The Proposed Multilayer Processing System

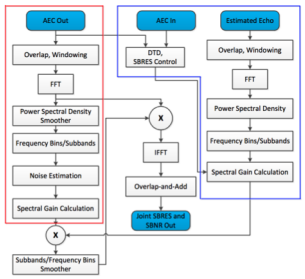

Fig. 2 The Proposed Joint SBRES and SBNR Alg.

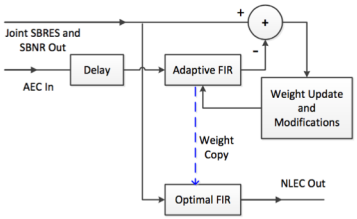

Fig. 3 The Proposed NLEC Alg.

- SBRES and SBNR: Subband spectral filtering approach
- The Freq. Bins/Subbands converter: Based on auditory critical bands
- Smoother: FIR median filter over freq.
- Noise Estimation: Search minimum statistics across band and time, no need VAD
- Double-Talk-Detection: Cross-correlation and ERLE based approaches

Noise Reduction and Echo Suppression Performance in VoIP

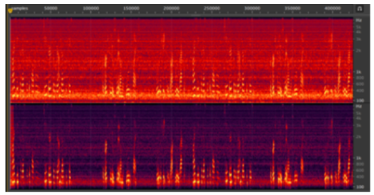

Fig. 4 Spectrograms. SBNR reduces ~19.3 dB noise (top = Off, bottom = On)

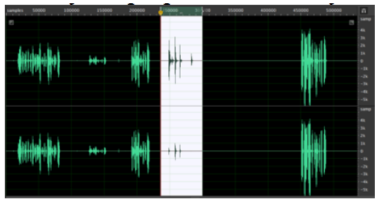

Fig. 5 Waveforms. SBRES reduces ~11.64 dB echo (top = Off, bottom = On)

Fig. 6 Waveforms. NLEC reduces ~35 dB echo (top = Off, bottom = On)

Fig. 7 Waveforms. SBRES+SBNR+NLEC reduces ~40 dB echo (top = Off, bottom = On)

Word-Error-Rate (WER) Performance in ASR

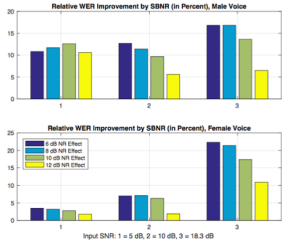

Fig. 8 Relative WER Improvement of SBNR Layer. There are 6,000 utterances for each type of noise (top = male, bottom = female)

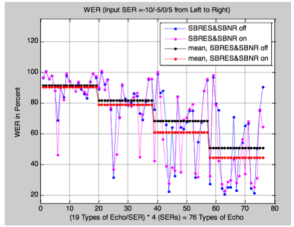

Fig. 9 WER of SBRES and SBNR Layers for 19*1486 = 28,234 words (the lower WER value is of better performance)

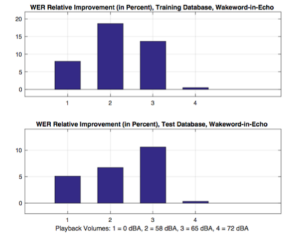

Fig. 10 Relative WER Improvement of NLEC Layer. There are 10,000 wake-words for each playback volume.

Features of the Proposed SBRES, SBNR, and NLEC Alg.

- SBRES and SBNR have been integrated ->greatly reduce MIPS and simultaneously suppress both echo and noise in a much more efficient way
- SBRES, SBNR, and NLEC are convergent fast
- Accurate suppression of noise and residual echo
- Echo and noise suppression amounts are adjustable
- Good voice quality, full-duplex, and ASR performances
- Can serve as efficient voice enhancement tool for many audio/voice related applications and devices when echo and noise are becoming complex and mixed.