DEEP LEARNING BASED AUTOMATIC VOLUME CONTROL AND LIMITER SYSTEM

Jun Yang, Philip Hilmes, Brian Adair, and David W. Krueger
Amazon Lab126, Sunnyvale, CA 94089, USA

Overview

- Automatic Speech Recognition (ASR) enabled smart speaker is the killer application.
- Volume Control (VC) is a key part in ASR based smart speakers.
- Existing VC methods are designed independently and hence result in poor audio quality and ASR performance.
- Driven by audio contents, a novel automatic VC (AVC) and limiter algorithm is proposed.
- Trained by wake-up word (WW) model in deep neural network (DNN) machine learning platform.
- Integrated with acoustic echo cancellation (AEC).
- Adaptively learns and tracks an effective signal level at the speed corresponding to the width of transient sound.
- No peak will go over the predetermined peak threshold. No clipping and harmonic distortion.
- Optimal ASR performance and audio quality can be achieved.

The Proposed AVC and Limiter Algorithm

- Look-ahead Buffer: a circular buffer to store a block of audio samples
- Time Constant Determination: audio content-driven approach
- Signal Event/Silence Detection: SNR-based Approach
- Signal Level Estimation:
  \[ G(n) = \frac{1}{N} \sum_{i=0}^{N-1} \left( x(n, i) \right)^2 \]
  \[ S(n) = S(n-1) + \xi \left( G(n) - S(n-1) \right) \]
  where \( \xi \) is attack time if \( G(n) > S(n-1) \); otherwise, \( \xi \) is release time.

Audio Performance

- No audible breathing artifacts, more dynamic range, more natural listening experience and balanced sonic experience than traditional AVC and Limiter.

Wake-up Word Recognition Performance

Correct Rate of Wake-up Word Detection versus Playback Volume. More green bars represent better ASR performance.

Features of the Proposed AVC and Limiter Alg.

- Audio content deep learning and data-driven features,
- Performs WW model training and statistic metric calculation for each audio feature,
- Clipping-free, very low latency,
- Very low MIPS,
- Natural listening and balanced sonic experience,
- No audible volume fluctuation during user’s adjusting volume,
- Can serve as efficient post-processor for many audio/voice related applications and device.