DEEP LEARNING BASED AUTOMATIC VOLUME CONTROL AND LIMITER SYSTEM

Jun Yang (IEEE Senior Member), Philip Hilmes, Brian Adair, David W. Krueger

Amazon Lab126, Sunnyvale, CA 94089, USA
Email: {junyang, philmes, briadair, dkkruege}@amazon.com

ABSTRACT

Automatic speech recognition is now playing an important role in volume control and adjustment of modern smart speakers. According to the recognition results by using the advanced deep neural network technology, this paper proposes an efficient processing system for automatic volume control (AVC) and limiter. The theoretical analyses, subjective and objective testing results show that the proposed processing system can offer a significant improvement for speech recognition performance during audio playback and improvement for audio playback performance in smart speakers. Driven by input data and audio contents, the proposed AVC is able to adaptively learn and track an effective signal level at the speed corresponding to the width of transient sound; the adaptation is frozen in the case of silence and noise periods. The proposed limiter measures the peaks and can guarantee that no peak will go over the predetermined peak threshold so as to avoid clipping and harmonic distortions.

Index Terms — Automatic volume control, limiter, wake-up word recognition, audio performance, clipping

1. INTRODUCTION

Automatic speech recognition (ASR) has found a very powerful use in volume control and adjustment of the modern smart speakers and hearing devices where such a device can generate an output audio signal with the desired sound pressure level (SPL) through applying a properly stable gain to the audio signal. These processing techniques are referred to as audio content based AVC and limiter (LIM). When users lower (or increase) the volume gain to less (or larger) than 0 dB, an AVC device can output the audio with the desired decreased (or increased) SPL. The applied actual gain could be different from the volume gain when volume gain is larger than 0 dB so as to avoid audible distortions due to clipping of peak signals.

As a matter of fact, it does not matter if users increase or decrease the volume gain; the desirable AVC processing should generate the output audio signals without audible volume fluctuations during users' adjustment of the volume or users' issuing of voice volume commands.

In addition, in some applications such as acoustic echo cancellation (AEC) system, the software solutions of AVC and LIM are very important due to the following three reasons: (1). AEC has full knowledge of the playback signal so that changing the volume gain does not result in an echo path change, (2). it is easy to configure the mapping between user volume changes and gain application, (3). AVC and AEC can be easily and tightly integrated so as to simplify the playback architecture and improve AEC performance.

More importantly, an AVC system should be not only responsible for controlling the volume for the device, but also responsible for limiting any signals that might exceed 0 dB digital full scale, i.e. preventing audio from clipping or saturation, which is one major cause of audible distortion. This suggests that a desirable AVC system needs to contain both AVC processing and limiter processing.

Although some AVC and LIM related systems have been proposed [1-6], these existing systems have significant drawbacks mainly because of generating either breathing, pumping, or distortion. Moreover, as shown in Section 3, the design of these existing systems is independent from the connected AEC processing and from the training for the speech recognition, which in turn cannot maximize the ASR performance during audio playback. The above problems prevent these existing AVC systems from practical use and being accepted by the users. It is the goal for this paper to propose a new AVC and LIM processing system that overcomes the above drawbacks so as to achieve the optimum processing performance. The proposed AVC and LIM system has been trained by wake-up word model on deep neural network (DNN) machine learning (ML) platform.

The rest of this paper is organized into the following three sections. In Section 2, we will present the proposed AVC system with emphasis on its audio content deep learning feature and open-loop processing mode. By using various testing results, Section 3 mainly shows that the proposed system can improve ASR keyword spotting (a.k.a. the wakeword or wake-up word) performance during audio playback, and can provide the users with a more natural listening experience and balanced sonic experience without noticeable volume fluctuations when users adjust the volume-knob or volume-button. Section 4 will make some conclusions and also make some further discussions.

2. THE PROPOSED AVC AND LIM SYSTEM

As mentioned in the previous section, the processing system consists of the AVC and LIM processing components. From an audio signal flow point of view, the AVC part is placed before the LIM. The LIM part should serve as the last step of the entire processing system so as to prevent output audio from clipping. However, in the AEC system, the output of LIM is sent to the AEC as the AEC reference signal.

2.1 The Proposed AVC Processing Algorithm

The proposed data-driven open-loop AVC algorithm with deep learning feature is shown as in Figure 1. This system mainly consists of six parts, i.e., signal content event and silence detection (i.e., tuned by deep learning), look-ahead buffering, time constant determination (i.e., tracking speed determination by data-driven), signal level estimation, frame gain estimation and learning, and final gain smoother.

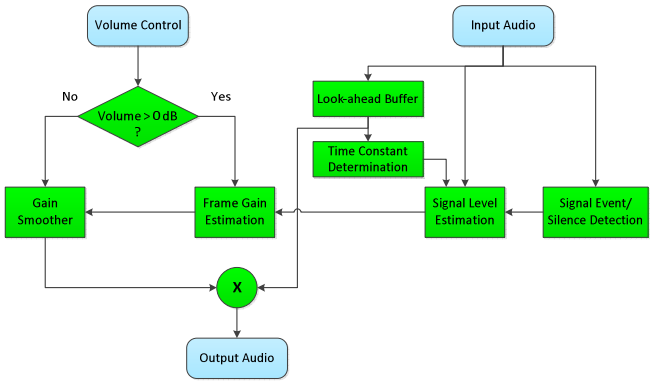

Figure 1 The Proposed Data-Driven Open-Loop AVC Alg.

In Figure 1, the block of "Look-ahead Buffer" is a circular buffer which stores a block of previous consecutive audio samples. The block of "time constant determination" (i.e., tracking speed determination driven by the related audio data) is to learn and obtain the appropriate attack and release times for different types of audio content so as to eliminate the audible artifacts. The principal feature of this processing block is to estimate the duration of transient sound by using the block of audio samples in the look-ahead buffer. The look-ahead buffer is about 35 ms. If the estimated duration is very short (e.g., less than 16 ms), then fast time constants (such as, 3 ms as the attack time and 30 ms as the release time) will be used for the "Signal Level Estimation" block, otherwise, the slow time constants (such as, 10 ms as the attack time and 100 ms as the release time) will be used for the "Signal Level Estimation" block. A more robust result can be obtained by using more layers and parameters.

The "Audio Signal Event and Silence Detection" block in Figure 1 is based on signal-to-noise ratio (SNR) and consists of 4 blocks, i.e., Frame Energy, Envelope Estimation, Floor Estimation, and SNR-based Comparison.

The "Frame Energy" block is implemented as follows.

$$G(n) = \frac{1}{N}\sum_{i=0}^{N-1}(x(n,i))^2 \qquad (1)$$

where N is the frame length in number of audio samples. The $x(n, i)$ is the i-th audio sample in the n-th frame. Therefore, $x(n, 0)$, $x(n, 1)$, ⋯, $x(n, N-1)$ are the block of audio signal in the n-th frame. The "Envelope Estimation" block is implemented by the following learning rule.

$$E(n) = E(n-1) + \beta(G(n) - E(n-1)) \qquad (2)$$

where β is a smoothing factor between 0.0 and 1.0. The "Floor Estimation" block is implemented according to the following learning rule:

$$F(n) = F(n-1) + \lambda(E(n) - F(n-1)) \qquad (3)$$

where λ is a smoothing factor between 0.0 and 1.0. The Comparison block is implemented as

If $E(n) > (\mu * F(n))$, then $K(n) = True$.

Otherwise, $K(n) = False$ (4)

The parameter μ is a SNR threshold which is an adjustable constant. The variable $K(n)$ represents audio event flag.

The "Signal Level Estimation" block of Figure 1 is implemented by a fast-attack and slow-release learning filter, that is, if the audio event flag $K(n)$ is false, then

$$S(n) = S(n-1) \qquad (5)$$

otherwise,

$$S(n) = S(n-1) + \xi(G(n) - S(n-1)) \qquad (6)$$

where if $G(n) > S(n-1)$, then $\xi = \eta_a$; otherwise, $\xi = \eta_r$. The parameters η_a and η_r are related to the attack and release time constants, respectively. They are determined by the processing block of "time constant determination" described above so that the attack and release time constants can match with the audio contents based on training with DNN ML platform.

The "Frame Gain Estimation" block of Figure 1 is implemented as follow.

If (S(n) * V(n)) > γ, then $p(n) = \gamma / S(n)$

otherwise, $p(n) = V(n)$ (7)

The parameter γ in Eq. (7) is a threshold which is an adjustable constant. The variable $V(n)$ is the volume gain adjusted by users.

The "Gain Smoother" block in Figure 1 is used to reduce the variation of the gain. The final gain is

$$g(n) = \alpha * p(n) + (1-\alpha) * g(n-1) \qquad (8)$$

where the factor α has the value between 0.0 and 1.0.

Multiplying the delayed block of input signals (i.e., the audio samples in the look-ahead buffer) by the obtained gain $g(n)$ in a way of sample-by-sample results in the output level of Figure 1 being well controlled smoothly.

It should be mentioned that wake-up word model training and statistic metrics are calculated for each of the above features. That means that each feature is trained on a machine learning model with large amounts of data. The statistical measures (such as mean, standard deviation, confidence level, etc.) are calculated for each feature model. The feature model with high confidence level (such as 95%) is then used in the proposed AVC system.

2.2 The Proposed LIM Algorithm

The proposed LIM algorithm creates a gain controlled signal shown in Figure 2 on the basis of the peak value of the input audio signal. The audio signal is physically delayed by an amount of time (i.e., look-ahead time). Once the control signal is ready to implement level adjustment, the audio is then sent ahead to the control element at the exact moment that the control signal arrives so as to make the adjustment. The proposed LIM not only meets the requirement of clipping-free and very low latency but also can be used to help any other audio processing, such as volume control, automatic gain control, 3D audio enhancement, and AEC, to prevent audio from clipping.

Figure 2 depicts the LIM gain curve corresponding to the three states if linear interpolation approach is adopted for both look-ahead state and release state.

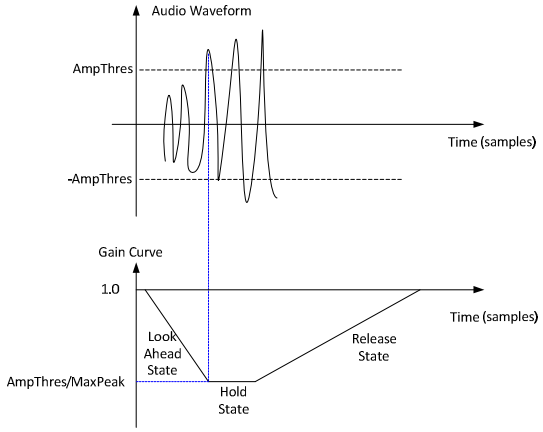

Figure 2 LIM Gain Curve for Linear Interpolation Approach

3. EVALUATIONS

In this section, we will present evaluation results and testing analyses of the proposed system in terms of speech recognition performance and audio quality.

3.1 Wake-up Word Recognition Performance

Test results are shown in Figure 3 through Figure 6, where the vertical axis is the "Correct Rate of Wake-up Word Detection", the horizontal axis is the "Playback Volume". More green bars represent better wake-up word detection performance. Figures 3 and 4 are for the case of 3 feet distance between near-end talker and the device-under-test, Figures 5 and 6 are for 6 feet distance. Obviously, the proposed AVC and LIM system has improved both the AEC performance and the wake-up word recognition rate.

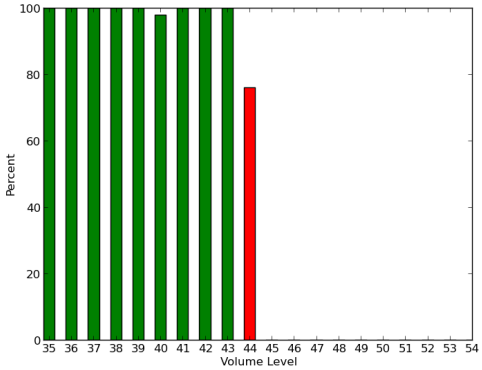

Figure 3. Correct Rate of Wake-up Word Detection versus Playback Volume from Traditional AVC and LIM (3ft).

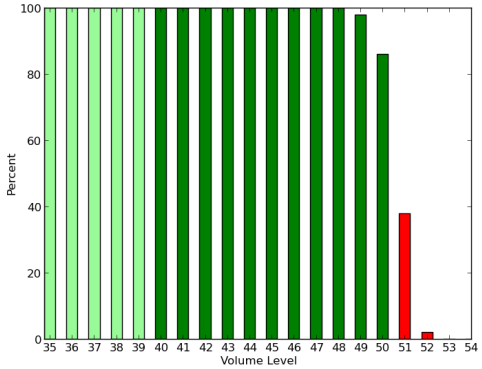

Figure 4. Correct Rate of Wake-up Word Detection versus Playback Volume from the Proposed AVC and LIM (3ft)

At 3 feet, wake-up word detection of traditional AVC and LIM approach drops to 76% at volume 44, then to 0% at volume 45 and greater. For the proposed AVC and LIM approach, wake-up word detection drops to 38% at volume 51, then to 0% at volume 53 and greater.

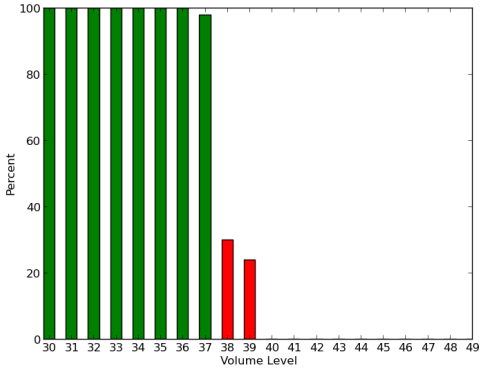

Figure 5. Correct Rate of Wake-up Word Detection versus Playback Volume from Traditional AVC and LIM (6ft)

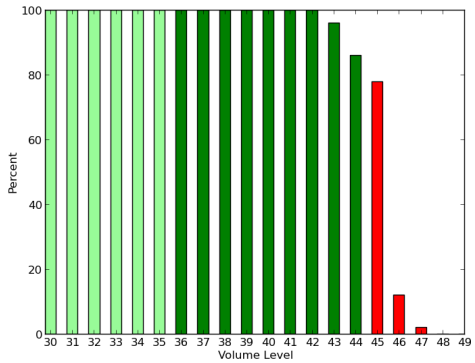

Figure 6. Correct Rate of Wake-up Word Detection versus Playback Volume from the Proposed AVC and LIM (6ft)

For the case of 6 feet, wake-up word detection of traditional AVC and LIM approach drops to 30% at volume 38, then to 0% at volume 40 and greater. For the proposed AVC and LIM approach, wake-up word detection drops to 78% at volume 45, then to 0% at volume 48 and greater.

3.2 Audio Performance

A representative example of breathing artifacts of traditional AVC and LIM is shown in the highlighted tail in Figure 8, while the highlighted tails in raw audio (Figure 7) and our proposed AVC and LIM output (Figure 9) are smoothly decayed without audible breathing artifacts. Also, Figure 9 shows more dynamics than Figure 8. Therefore, the proposed AVC and LIM can provide users with a more natural listening experience and balanced sonic experience without audible volume fluctuations during user's adjusting volume-knob, volume-button, volume-ring, voice UI volume commands, or GUI volume commands.

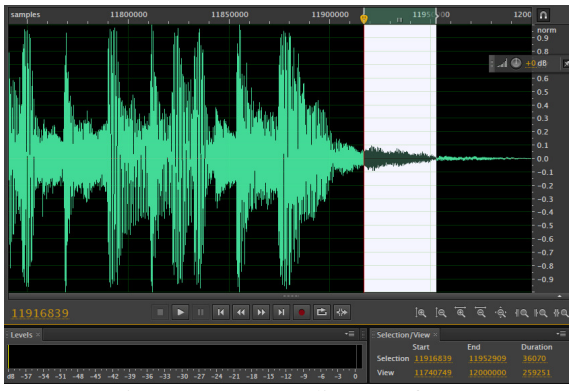

Figure 7. Input Audio Waveform

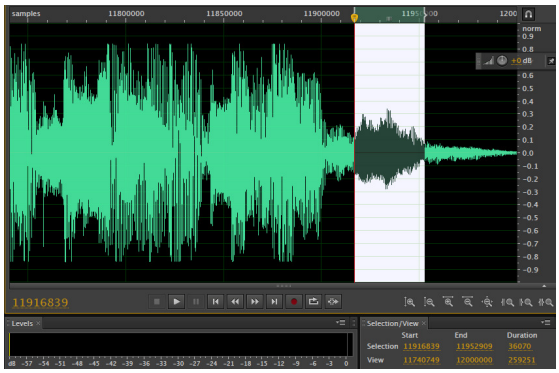

Figure 8. Output Waveform Processed by Traditional AVC and Limiter for 24 dB Input Volume

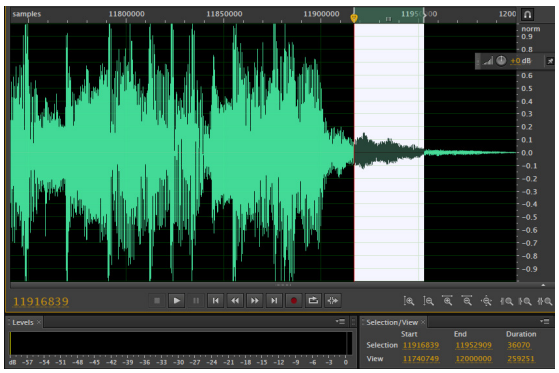

Figure 9. Output Waveform Processed by our Proposed AVC and Limiter for 24 dB Input Volume

4. SUMMARY

As we pointed out in Section 1, traditional AVC and LIM systems are not of data-driven working principles where tracking signal level is at a constant speed. The design of these traditional AVC and LIM ones is independent of the AEC system and also independent of training for the speech recognition, which results in significant drawbacks.

This paper presents a new AVC and LIM system on the basis of audio content deep learning and data-driven features in an open-loop processing mode. Also, the proposed system performs the wake-up word model training and statistic metrics calculations for each audio feature in a more efficient way. More importantly, the proposed limiter processing provides clipping-free output and is of very low latency. The proposed AVC and LIM system has been efficiently implemented in a hardware computing platform and extensively evaluated in both real-time acoustical objective testing and subjective listening tests. Due to its simplicity in computational complexity, the MIPS requirement incurred by the proposed system is also very small. All of the above shows that the proposed system can serve as a very efficient AVC and limiter processing tool for many audio/voice related applications and devices.

5. REFERENCES

[1] Dimitrios Giannoulis, Michael Massberg, and Joshua D. Reiss, "Digital Dynamic Range Compressor Design – A Tutorial and Analysis," *J. Audio Eng. Soc.*, Vol. 60, No. 6, June 2012, PP. 399 - 408

[2] Jicai Liang, Song Gao, Yi Li, "Research on Dynamic Range Control used to Audio Directional System," *2011 Internal Conference on Mechatronic Science, Electric Engineering and Computer*, August 19-22, 2011, PP. 498 - 501, Jilin, China.

[3] R. J. Cassidy, "Level Detection Tunings and Techniques for the Dynamic Range Compression of Audio Signals," *The 117th Convention of the Audio Engineering Society*, Oct. 2004, convention paper 6235, San Francisco, CA, USA.

[4] Lucio F. C. Pessoa, Mao Zeng, "Automatic Level Control Device for Digital telephony Systems," *Global Signal Processing Expo and Conference* (GSPx), Sept. 27-30, 2004, Santa Clara, CA, USA

[5] M. A. Stone, B. C. J. Moore, "Effect of the Speed of a Single-Channel Dynamic Range Compressor on Intelligibility in a Competing Speech Task," *J. Acoust. Soc. Am.*, vol. 114, 2003, PP. 1023 - 1034

[6] P. Hamalainen, "Smoothing of the Control Signal without Clipped Output in Digital Peak Limiters," *2002 International Conference on Digital Audio Effects (DAFx)*, PP. 195 – 198, Hamburg, Germany