

Wave Arts VQE

Bill Gardner
Wave Arts, Inc.
Nov 16, 2011

Introduction

Wave Arts VQE (Voice Quality Enhancement) is a software library implementing acoustic echo cancellation for voice over IP (VoIP) applications. The principal features are:

- Acoustic echo cancellation (multi-band FRLS algorithm)
- Acoustic echo suppression (optional lo-fi method)
- Noise suppression
- Auto-gain control
- De-clipping (useful for headsets)
- Speaker distortion compensation
- Cross-platform audio API: Mac OS-X (CoreAudio), Win XP (MME), Win Vista (WASAPI), and Win 7 (WASAPI)
- Device selection, volume control, level metering, hot swapping support
- Sample rate conversion
- Device calibration to compensate for sampling rate offsets
- Supports 8 kHz (narrowband) and 16 kHz (wideband) sampling rates
- Standalone demo application on Mac and Win
- Integration code for PJSIP

An application integrating with VQE would use the VQE API to open audio devices and stream audio, then all echo cancellation and other signal processing is built-in. It's cross-platform (Windows and Mac OS-X for now).

If you are using PJSIP, there is integration code that implements a PJSIP audio device using VQE, you need only enable the VQE audio driver in the PJSIP configuration.

There is also a standalone VQE demo application with JUICE user-interface which demonstrates VQE and provides additional example code. The demo application plays some speech that asks you questions, records your answers, and then plays back the echo canceled recording of your answers. The application has a built-in user guide.

Wave Arts VQE is released under a dual GPL/commercial license. Developers can download the source and test drive, or use it in GPL open source apps, and if you want to use it in a closed source application you must purchase a commercial license.

VQE Signal Processing

This section will describe the signal processing features of VQE, which include echo cancellation, echo suppression, noise suppression, sample rate offset calibration and compensation, auto gain control, and de-clipping.

About Echo Cancellation

When sound from the remote party is played on the speaker at your location, it is picked up by your microphone and sent back to the remote party. They will hear their own voice delayed by the roundtrip transmission time, which is typically a few tenths of a second. Because of the delay the echo can be quite annoying to the talker. An echo canceller removes the echo from the recorded signal so the remote party hears only your voice.

Headsets with integrated boom microphones are useful for VoIP application because they generate very little echo; the sound from the headset that reaches the microphone is much lower than sound from your voice. When using speakers and desktop microphones, a great deal of the playback sound is picked up by the microphone. On laptops, where the microphone is close to the speakers, the echo from the speaker is actually much louder than your voice. So a good echo canceller is essential when using speakers and microphones.

This document will use a lot of echo cancellation terminology. The remote party is called the "far" side, as opposed to the "near" side at your location. When only one side is speaking at a time, this is called "single-talk". When both sides speak at the same time, this is called "double-talk". A good echo canceller will permit double-talk, i.e., allow both sides to speak at the same time, but some only allow one party to speak at a time.

Strictly speaking, the term "echo cancellation" refers to cancelling, or subtracting, the echo from the recorded signal. An adaptive digital filter processes the signal from the far side and subtracts the result from the near side recording. The filter coefficients are adapted to minimize the resulting error, hence the filter is adapted to reduce the recorded echo. The filter is controlled so it adapts only when the far side is talking; the adaptation is stopped when the near side is talking to prevent it from attempting to cancel your voice.

The electroacoustic signal gain from speaker to microphone is called "Echo return loss" (ERL). The amount of attenuation provided by the adaptive filter is called "Echo return loss enhancement" (ERLE).

The term "echo suppression" refers to attenuating the recorded signal when it contains echo. Similar to an echo suppressor, a "noise suppressor" attenuates background noise. Noise suppressors are also used along with echo cancellers to attenuate uncanceled echoes. These are also called "non-linear processors" (NLP). The basic idea is that low level signals are muted.

VQE contains elements of echo cancellation, echo suppression, and noise suppression algorithms. It would be best described as an "echo control" (EC) system, but we simply call it an echo canceller.

VQE Echo Cancellation

The VQE echo canceller contains an adaptive filter to cancel echoes and an echo/noise suppressor to mute background noise and uncanceled echoes. The adaptive filter is implemented using a multiple band fast recursive least squares (FRLS) algorithm. The input signal is divided into different frequency bands, and in each frequency band a FRLS echo canceller is applied. The FRLS algorithm was chosen because it converges very quickly. While the FRLS algorithm is computationally expensive, the multi-band approach allows it to be deployed efficiently because the individual bands run at a lower sampling rate.

The FRLS algorithm uses a “two-path model” to control adaptation. Two filters are run at all times; a background filter which is always adapting, and a static foreground filter which is actually used to cancel echoes. The errors from the filters are periodically compared, and when the performance of the background filter exceeds that of the foreground filter, the filter coefficients are uploaded from the background filter to the foreground filter.

The suppressor portion of the echo canceller is implemented using a noise gate which opens when the near side is talking. The near talk detector uses adaptive estimates of the recording's noise floor and signal level to determine an activity threshold. The gate opens when the threshold is exceeded. Similar logic is used to determine far side voice activity. During far side activity, there is more complicated double-talk logic that controls the suppressor. Double talk detection uses estimates of the ERL and ERLE to determine the expected level of the recorded echo after cancellation. If the actual recorded signal exceeds this by a threshold amount, the suppressor asserts the double-talk condition and opens the gate. This algorithm is simpler and incurs less latency than double-talk detection based on signal coherence measurements.

VQE also contains a strictly suppressive algorithm called “Acoustic echo suppressor”. This algorithm allows only one side to speak at a time, the other side is muted.

Sample rate offset

On a desktop computer, the speakers and microphone will typically be different audio devices each with its own sampling clock. Even when both devices are configured to run at say 16000 Hz, the actual rate will differ slightly due to the accuracy of the clocks; this is called "sample rate offset". Typical sampling rate offsets due to clock accuracy are less than 50 ppm (parts per million). However, when using Windows MME drivers the sampling rate offset can be greater than 6000 ppm, which is perhaps due to errors in internal sample rate conversion software. Because audio drivers may do internal sample rate conversion, there is always the possibility of unexpected sampling rate errors, and these errors can be much larger than expected from clock differences.

The adaptive filters used in echo cancellers expect the sampling rates to be exactly the same, even a small offset will reduce the amount of echo cancellation achievable. Also, sampling rate offsets require the adaptive filter to constantly re-adapt to a forever changing target. When the far side is quiet and then begins speaking again, the filter coefficients from previous adaptation will be out of date and will not cancel well until they are updated anew. Sample rate offset also causes sample buffering problems. Echo cancellers process equal numbers of recorded (near

side) and playback (far side) samples, so the recording buffers will accumulate extra samples when the record rate exceeds the playback rate, and the recording buffers will lose samples when the record rate is less than the playback rate. Eventually record buffers must be discarded or replicated which causes a discontinuous jump in system latency.

VQE has several mechanisms to address sample rate offset:

1. Upsampled devices using fast resampler
2. Calibration measurement of sample rate offset
3. Accurate sample rate conversion
4. Sample buffer synchronization
5. Echo suppression failover

As mentioned earlier, when using Windows MME drivers, the sampling rate offset can be extremely large, over 6000 ppm, even when using the same audio hardware for playback and recording. This seems to be due to internal sample rate conversion in the MME drivers. One way to avoid the offset is to run the audio device at the maximum sampling rate of 48 kHz. VQE provides a convenient upsampled device port which can open the recording and playback devices at an arbitrary sampling rate. The sample rate conversion is implemented using pre-computed coefficient tables so the implementation is efficient. In some cases running the devices at 48 kHz will give perfect synchronization while running at 16 kHz or 8 kHz will have large offsets. In other cases, the upsampling will eliminate the large sample rate offset and yield a small < 50 ppm offset expected when using different devices.

To deal with the unavoidable sample rate offsets when using different hardware devices, VQE provides both a calibration system to measure the sample rate offset and sample rate conversion to compensate for the offset. The calibration system plays a sine tone, records it, and then calculates the sampling rate offset by measuring the frequency of the recorded tone. In quiet conditions the measurement is accurate within 1 ppm. After measurement, the sample rate offset is compensated by resampling the recorded audio.

VQE provides very accurate sample rate conversion which can resample by an arbitrary offset. For a given device pair, the calibration measurement need only be done once because the sample rate offset is very stable over time. The measured offset can be stored in a database and used to open the devices with the appropriate compensating resampling.

If the sample rate offset is not compensated, the input buffers will either grow or shrink over time. VQE uses an input buffer FIFO to gather recorded buffers, match them with playback buffers, and pass buffer pairs to the echo canceller. The FIFO contains logic to deal with underruns and overruns which guarantees that the echo canceller algorithm sees buffer pairs with a fixed latency relationship.

Finally, if the sample rate offset is not corrected by upsampling or calibration/resampling, the adaptive echo canceller filter will likely perform poorly and may not cancel echo at all. In this case the echo canceller algorithm works in purely suppressive mode.

Audio Latency

The latency (delay) of the desktop audio system is generally unknown. The length of the playback buffers serves as a lower bound on the delay, but the delay will be larger due to internal buffers. The acoustic delay from speaker to microphone is usually negligibly small (say a few milliseconds) compared to the buffer delay (typically 100-200 milliseconds). It is not essential that the delay be known, the adaptive filter will simply converge to a response that starts with a number of zeroes corresponding to the delay. However, if the adaptive filter is computationally expensive it is best to align it so that it does not compute a lot of leading zeros. This is done by delaying the far side signal going into the echo canceller so that the far signal is aligned with the corresponding echo return from the microphone. Knowing the system latency exactly is also helpful for doing double-talk detection.

VQE uses a calibration system to measure the audio latency of the currently selected device pair. A short beep is played when the devices are opened, and the initial recording is analyzed using a matched filter to determine the delay of the beep. A confidence estimate is also made using the signal to noise ratio of the recording; the recorded estimate is used only when there is a quiet recording, otherwise the playback buffer sizes determine a lower bound on the latency.

Speaker distortion compensation

The small loudspeakers used in laptops and desktop systems distort easily, particularly when reproducing low frequency sounds. Distortion causes the creation of tones that are not in the original sound. The distortion products can't be cancelled by the adaptive filter. In order to reduce the amount of distortion it suffices to reduce the level of low frequency sounds by applying a bass reduction filter. For high fidelity speakers, no reduction is needed. For midsize desktop speakers, a moderate amount of reduction is applied. For tiny speakers used in laptops, a large reduction is applied. VQE contains EQ filters that apply the bass reduction to the playback signal.

Auto-gain

VQE provides a simple auto-gain control (AGC) algorithm which will change the recording device volume in response to detected signal levels. Low detected levels cause the gain to be increased, and high levels or clipped levels cause the gain to be reduced. Once the gain is reduced, further gain increases are disabled. The motivation behind the AGC is to prevent having the volume set so low that the talker is hard to hear, and also to prevent clipped levels that can cause distortion.

Unlike a digital signal compressor, the AGC changes the analog volume control of the audio device. When the signal level is too low, the recording may contain a lot of quantization noise, so digitally amplifying the signal would also amplify the noise. When the signal is loud enough to exceed the range of the recording ADC, the resulting clipping distortion will not be affected by lowering the digital gain. Hence it is important for the AGC to manipulate the analog volume control.

De-clipping

In cases where the recorded signal exceeds the range of the ADC, i.e., before the AGC has a chance to lower the gain, the recording will contain clipping distortion. This is possible when speaking loudly into a headset microphone. The De-clipping algorithm automatically softens the distortion by applying a smoothing lowpass filter to the clipped audio.