The shoulds and shouldn’ts of providing intelligible speech through VQE

The Fundamentals of Voice Quality Enhancement

By Scott Kurtz

Listen up! You can deploy voice quality enhancement (VQE) techniques to improve speech intelligibility and reduce listening effort—in other words, to deliver much better speech quality—in the telecommunications systems you design. The main ways to achieve VQE: remove unwanted echo and noise and adjust the listening level of the signal to one that’s comfortable for the user. Furthermore, besides using VQE building blocks to enhance speech quality, you can apply the same technology to other speech-related applications—for example, speech recording and recognition.

VQE is becoming increasingly important because the telecommunications landscape continues to change and grow as the customer base yearns for new ways to communicate and because, believe it or not, more people throughout the world carry cell phones than pens. Much to the annoyance of some, droves of people now carry on conversations in trains, planes and automobiles; while walking on city streets and eating at restaurants; stretched out on the beach and sprawled in public parks. Everywhere you look, it seems, people are chatting away.

Aside from the social implications, those environments aren’t ideal for quiet, clear communication. At the same time, the telecommunications infrastructure itself is evolving—more voice traffic is moving from traditional circuit-switched networks to packet-switched nets.

Although technology has made these new communication options possible, in many ways it must improve to meet the new demands. As a telephone user, you’re accustomed to enjoying excellent speech quality when using the traditional telephone network. On the other hand, you’ve had to accept a far lower speech quality when using some of the newer technology, like cell phones. VQE is the umbrella that encompasses the technology intended to improve speech quality in both the newer and traditional telecommunications systems.

Let’s consider the phenomena that degrade speech quality, some methods you can use to fight back, and the best way you can integrate those methods into your telecommunications systems.

Impairments to Speech Quality

From a listener’s perspective, you can categorize perceived speech quality using two criteria—intelligibility and listening effort, or the amount of effort or concentration required to understand a speaker. The importance of intelligibility is obvious. As for listening effort, when listening you want to think about the content of your conversation without having to concentrate

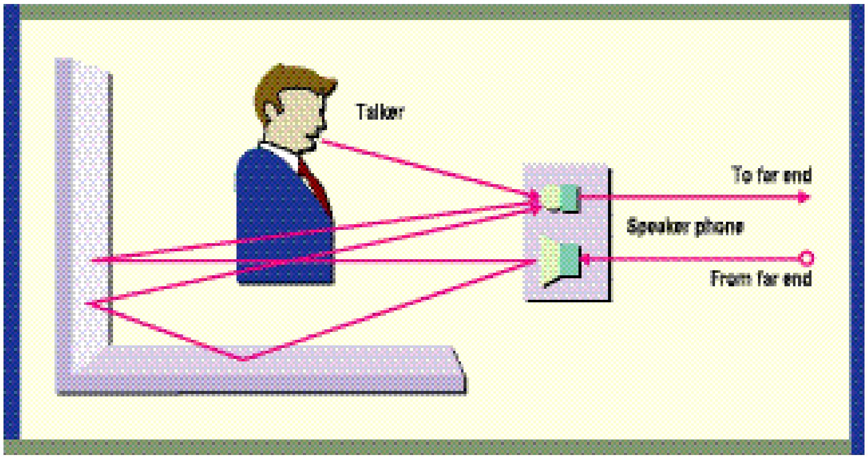


Figure 1: Hybrid circuits in the telephone network convert four-wire interfaces into two-wire interfaces. Although these circuits are designed to mitigate echo caused by coupling between the four-wire input and the two-wire output, they're far from perfect at doing so.

solely or primarily on understanding the words.

Here's a little aside: Jim James, the director of AT&T's Speech Quality Test Lab, provides an interesting business perspective. When discussing AT&T's motivation for achieving high-quality speech, he said: "The better the speech quality, the longer people will talk on the phone. The longer people talk on the phone, the higher the revenue."

To achieve high speech quality, you must first identify and understand the phenomena that reduce speech quality; then you can learn how to mitigate their effects. The main impairments are noise, variations in signal level (attenuation), and echo.

Ideally, you want to listen to a speaker without any background noise—a condition that seems unlikely in an age of rising use of cell phones in noisy environments. The types of background noise vary with the environment. In a vehicle, engine and road noise are significant. In a factory or construction environment, machinery can produce extremely high levels of noise. In an office, noise can come from office equipment, forced air heating, air conditioning, and other conversations. At home, the main causes are appliances, running water, audio/video gear, and A/C.

Making matters worse is that people don't all speak at the same volume and that voice signals can be attenuated at various points in the telephone network. Attenuation has two primary causes: loop resistance when the phone is far from the telephone central office or multiple phones are off-hook in parallel on the same circuit, and loss due to the distance between the speaker and the microphone when a speaker or hands-free phone is used. Consequently, the amount of attenuation will vary from speaker to speaker and circuit to circuit. The bottom line is that there's a range of signal levels that make listening most comfortable. If the level is too high or too low, understanding the person speaking becomes more difficult.

As for the third phenomenon, echo, you'll encounter two types in telecommunications systems—electrical and acoustical.

The first is caused by circuits,

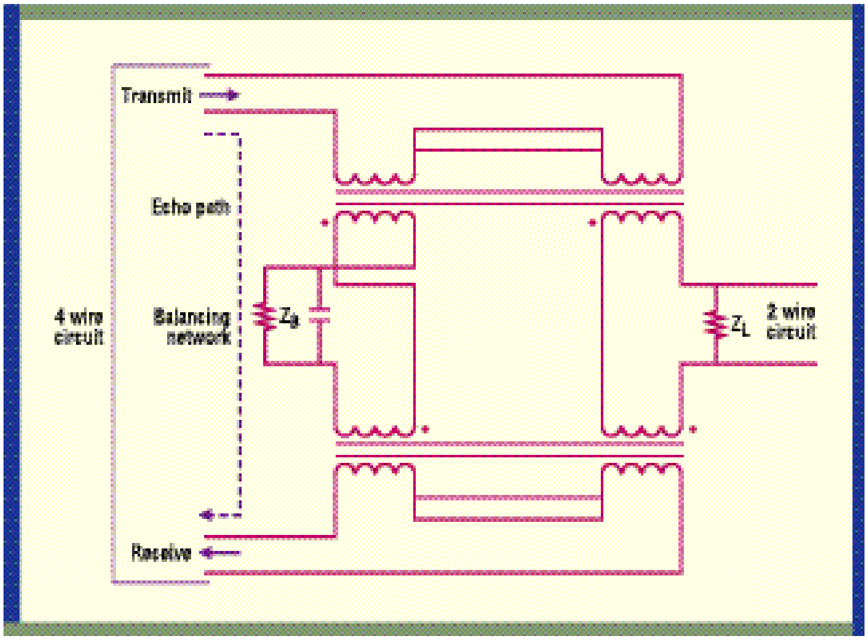


Figure 2: Acoustic echo is caused by direct and indirect feedback from speaker to microphone.

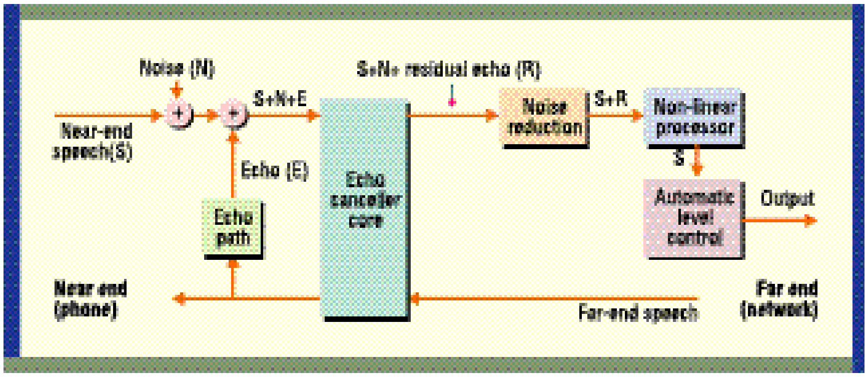


Figure 3: One configuration for integrating building blocks into a VQE solution. Figure 3 shows one configuration for integrating echo cancellation, noise reduction, and automatic level control into a VQE solution.

known as hybrids, that perform four-wire to two-wire conversion, shown in Figure 1. In the process of performing the conversion, some signal reflection occurs, and the person at the other end of the conversation perceives the reflection as echo.

Acoustic echo is caused by feedback between speaker and microphone in a telephone handset, conference phone, or hands-free phone, as shown in Figure 2. In addition to the direct feedback between speaker and microphone, the output of the microphone feeds back to the microphone after multiple reflections in, for example, an office, conference room, or automobile. Again, the person at the far end perceives the feedback as echo.

The amount of degradation in speech quality caused by echo is a function of both the level and the delay of the echo signal with respect to that of the original speech.

Enhancing Speech Quality

Although it's beyond the scope of this article to describe VQE algorithms in detail, it's helpful to understand the most important requirements that these algorithms must satisfy.

First, a good noise reduction algorithm should reduce background noise whatever the existing conditions, thereby increasing the signal-to-noise ratio of the speech signal. Second, the algorithm should have a minimum effect on speech signals. Third, it should be able to adapt to different conditions but also have programmable parameters for threshold noise level, below which noise reduction is disabled, and for maximum noise attenuation or some other measure of noise reduction aggressiveness.

At the same time, a good noise reduction algorithm shouldn't degrade speech quality, especially under conditions where there is little or no noise present. Nor should it introduce excessive delay or degrade the performance of fax or voiceband data modems.

The applications for noise reduction extend beyond telecommunications systems. Consequently, many classes of noise reduction algorithms exist. Two classes apply best to telecommunication applications: those that operate with a single microphone as the input and those that operate with an array of microphones as the input; the latter technique is also known as acoustic beam-forming. Of the two, the class for a single microphone is more generally applicable.

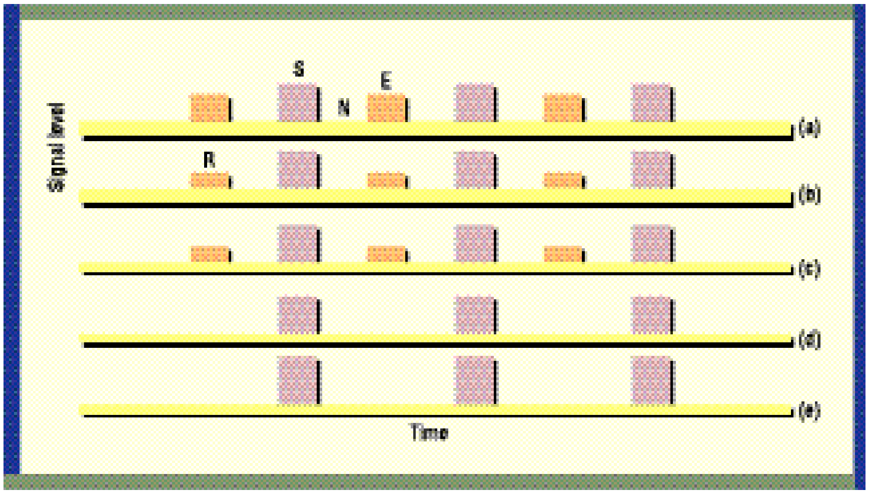


Figure 4: Each of the graphs displays signal level versus time at a different test point for the VQE solution in Figure 3. The echo canceller core receives a signal consisting of the near-end speech, noise, and echo (a) and reduces the level of the echo to form the residual echo (b). Similarly, the noise reduction algorithm eliminates the noise (c). The NLP then removes the residual echo (d). Finally, the ALC boosts the level of the near-end speech signal (e).

Single-microphone noise reduction is usually accomplished by decomposing the input signal into multiple frequency bands, applying appropriate noise reduction in each band, and recombining the bands. With acoustic beam-forming, it's assumed that the desired speech signal comes from one source at a time and that noise comes from multiple directions. Beam-forming allows you to direct the maximum gain of a microphone array in the direction of the desired signal source. Since the noise comes from directions where the gain is less, the net effect is an improvement in the signal-to-noise ratio. Furthermore, you can use the multiple microphones to estimate the noise and then subtract the noise estimate from the signal, thereby improving the signal-to-noise ratio even more.

When implementing any type of noise reduction, it's important to minimize the distortion of the desired speech signal. You must trade off the degree of noise reduction with distortion of the desired signal. Make sure that a good automatic level-control algorithm accompanies the one for noise reduction. At the least, this algorithm should provide gain or loss to a speech signal while speech is present, provide loss when the signal level is very high, provide gain when the signal is too low, disable itself in the presence of high-speed voiceband modems, and have programmable parameters for maximum gain and attenuation, target level, threshold speech level (below which gain is not applied), and the like.

When implementing any type of noise reduction, it's key to minimize the distortion of the desired speech signal.

Naturally, there are things a good automatic level-control algorithm shouldn't do, like not cause percep-

tible artifacts in the speech signal, amplify noise when speech isn't present, distort signaling tones (such as DTMF tones), or introduce excessive delay.

Just as there are two types of echo, electrical and acoustical, there are two types of echo cancellers—network (and line) and acoustic. As you may surmise, they're very similar, with their differences being attributable to the nature of the different types of echo.

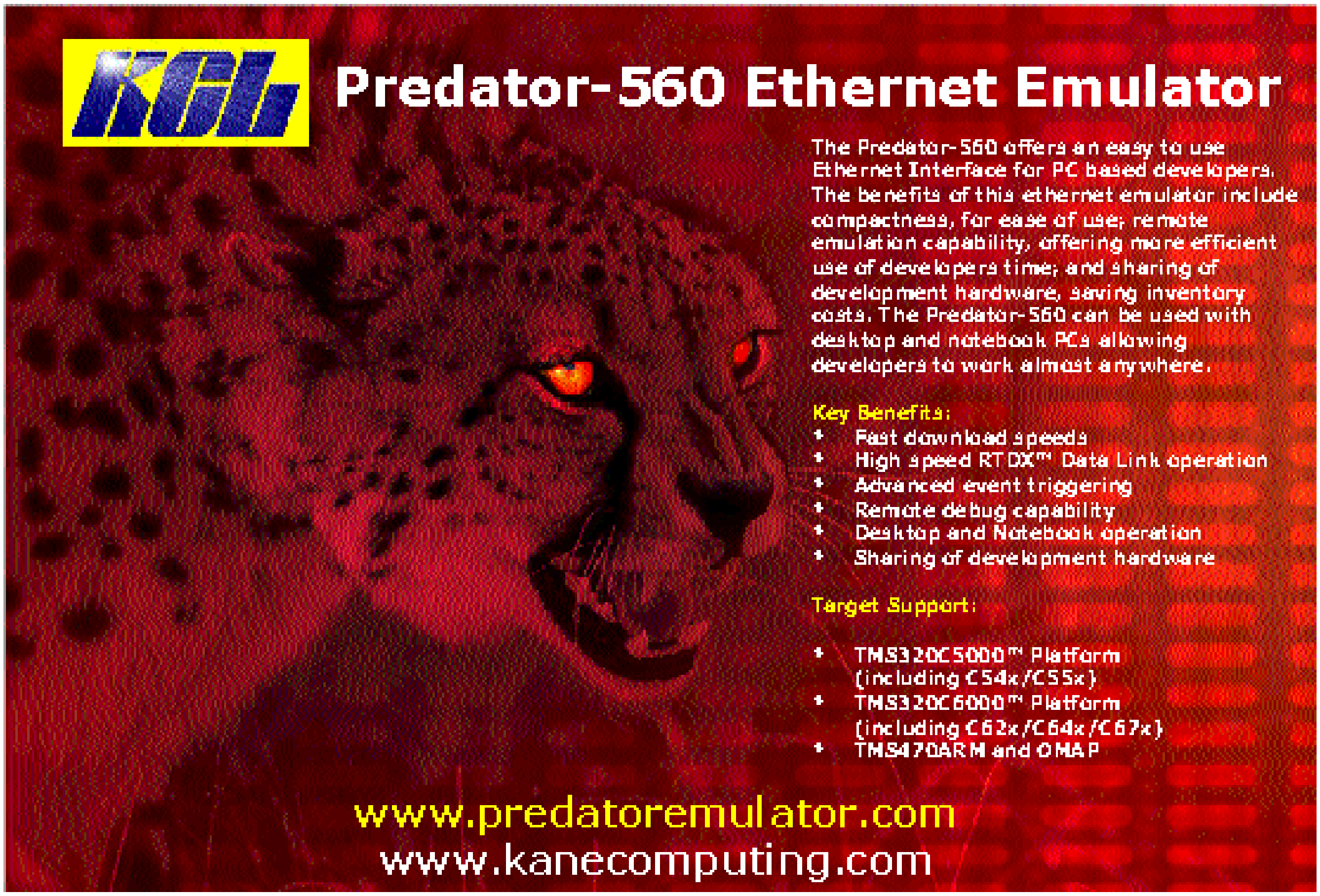
In general, a good echo canceller should meet six positive requirements and three negative ones.

First off, it should meet or exceed the requirements set forth in ITU recommendation G.168-2000 (network echo canceller) or G.167 (acoustic echo canceller). It should provide deep echo cancellation (high return loss enhancement) and have a robust double-talk detector to prevent divergence when both parties are speaking at the same time. Also, it should include nonlinear processing to remove residual echo and a comfort noise generator to prevent unwanted transitions between silence and background noise. In addition, it should disable cancellation in the presence of high-speed voiceband modems. On the negative side, a good canceller shouldn't diverge in the presence of tones, degrade the performance of fax or voiceband data modems (network error correction only), or introduce excessive delay.

The key to integrating a VQE solution is to put the various algorithms together in such a way that maximizes speech quality enhancement.

Figure 3 shows one possible configuration for enhancing the quality of the speech signal sent from the near end to the far end. Figure 4 depicts sample signals at various points in the diagram; the signal designations follow the same convention as that in Figure 3.

The echo canceller core is the first algorithm to operate on the near-end signal. The near-end input to the echo canceller consists of three components: near-end speech (S), additive noise (N), and added echo (E) of the far-end speech, as shown in Figure 4a. The job of the echo canceller core is to reduce the echo component from the signal.



The core removes most of the echo but leaves a small residual amount. It therefore puts out a composite signal containing near-end speech (S), additive noise (N), and residual echo (R), shown in Figure 4b.

The output of the core is fed to the noise reduction algorithm, which increases the signal-to-noise ratio of the signal by reducing the background noise level. The output of the noise reduction algorithm ideally consists of the near-end speech (S) plus residual echo (R), shown in Figure 4c.

The output of the noise canceller is then fed into a nonlinear processor (NLP). The NLP's function is to suppress the residual echo based on estimates of the signal level of the far-end speech and the residual echo. It's advantageous to place the noise reduction before the NLP so that the noise doesn't affect the estimate of the residual echo signal level. Furthermore, if the noise reduction were to follow the NLP, the NLP would sometimes suppress the background noise, as well as the residual echo. The resulting variations in noise level would reduce the performance of the noise canceller.

The output of the NLP, shown in Figure 4d, is fed to the automatic level-control (ALC) function. The ALC function introduces the appropriate amount of gain or loss in order to create an output speech signal whose signal level is appropriate for good perceived speech quality, as in Figure 4e. ◆

Scott Kurtz (scott.kurtz@adaptivedigital.com) is the vice president of engineering for Adaptive Digital Technologies, Inc. in Conshohocken, Pa., where he has focused on the development of the company's DSP-based products and technology. He has specialized in digital signal processing and digital communications over the course of his 18-year career, beginning at RCA's Government Communications Systems Division. At InterDigital Communications, he was instrumental in developing the first digital wireless local loop telephone system, which was the precursor to today's digital cellular telephone systems. He holds twelve patents, with two more pending.