

3

Voice Quality

3.1 Introduction

A common joke among IP telephony engineers is to say that if they had proposed to carry voice over IP a couple years ago, they would have been fired. This remains a private joke until you make your first IP phone call to someone on an old PC without a headset, to find out that the only person you heard was yourself (this is no longer true today ... even PCs have software echo cancelers). Another way to find out why there really is a problem with IP telephony is to try a simple game: “collaborative counting”.

Collaborative counting has a simple rule: if you hear the person you talk to say ‘ n ’, you immediately say ‘ $n + 1$ ’. In order to compare ‘classic’ telephony with IP telephony, you first make a regular phone call to someone you know and say ‘1’, he goes ‘2’, etc. Keep an eye on your watch and measure how long it takes to count to 25.

Then you make an IP phone call and play the same game. In all cases, it will take much longer ...

The problems we have just emphasized, echo and delay, have been well known to telephone network planners since the early days of telephony, and today’s telephone networks have been designed to keep these impairments imperceptible to most customers.

When carrying voice over IP, it becomes much more difficult to control echo, delay, and other degradations that may occur on a telephone line. As we will see, it will require state-of-the-art technology and optimization of all components to make the service acceptable to all customers.

However, once echo and delay are maintained within acceptable limits by proper network engineering, VoIP can use voice coders, such as G.722 (‘wide-band’ coder), which provide an absolute voice quality beyond that of current PSTN networks. It is frequently heard that VoIP can reach ‘toll quality’; in fact, in the future VoIP will provide a voice quality that exceeds ‘toll quality’ through rigorous planning and design.

3.2 Reference VoIP media path

The media path of IP telephony calls can be modeled as shown in Figure 3.1, when there is no POTS or ISDN terminal involved. The situation gets a little more complex when interworking with an ISDN phone through a gateway (Figure 3.2):

When the gateway interfaces with an analog network, the user–network interface is in most cases using only 2 wires (incoming and outgoing signals share the same pair), and a 4-wire/2-wire hybrid is required (Figure 3.3). The model includes the most significant sources of voice quality degradation:

- The IP network introduces packet loss, delay, and jitter.

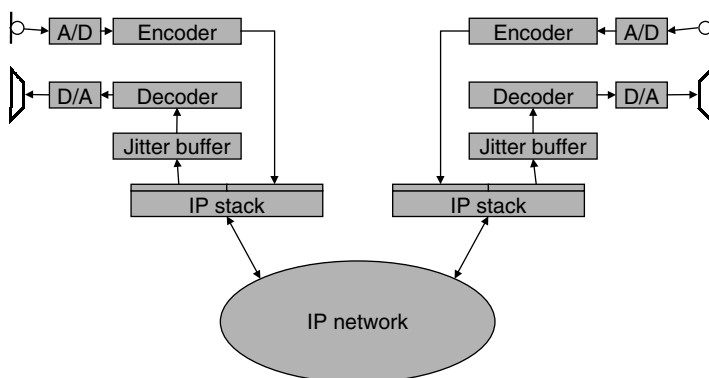


Figure 3.1 Reference VoIP media path.

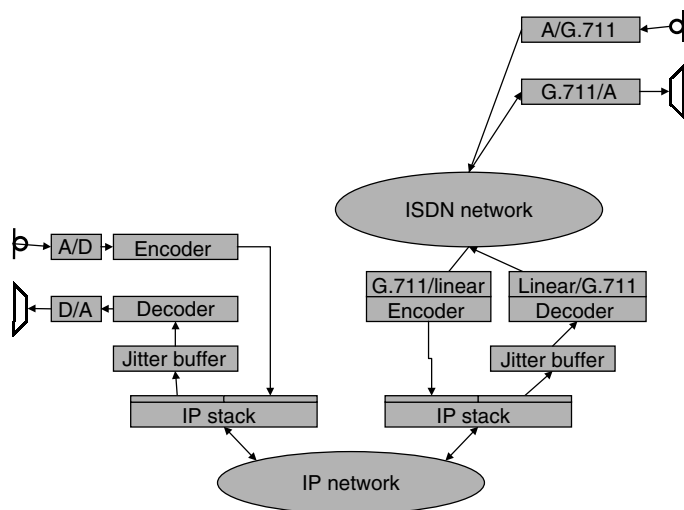


Figure 3.2 Reference VoIP to ISDN path.

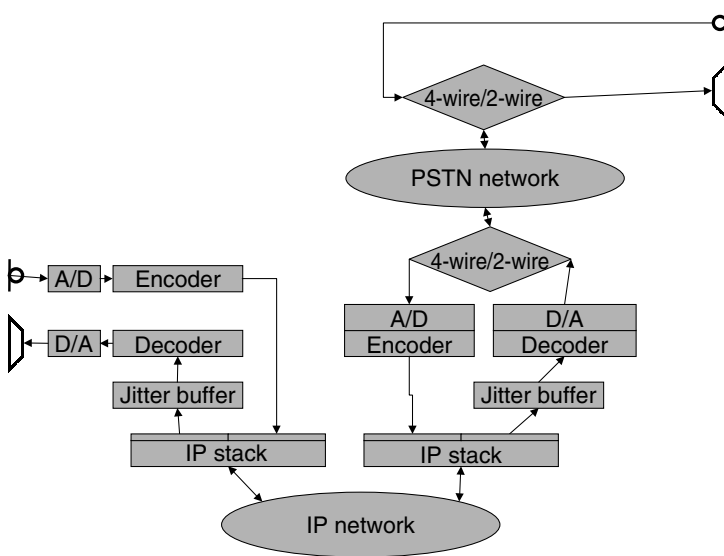


Figure 3.3 Reference VoIP to POTS path.

- The jitter buffers (JBs) influence end-to-end delay and frame loss.
- The acoustic interfaces introduce acoustic echo.
- The 2-wire/4-wire interfaces introduce electric echo.
- The PSTN network, which potentially introduces further delays.

In this chapter we describe the main factors influencing end-user perception of voice quality. Most of those factors are common to switched circuit telephony and IP telephony. However, IP telephony has some unique characteristics, such as long delays, jitter, and packet loss, and therefore requires a new framework for assessing voice quality.

3.3 Echo in a telephone network

3.3.1 Talker echo, listener echo

The most important echo is **talker echo**, the perception by the talker of his own voice but delayed. It can be caused by electric (**hybrid**) echo or acoustic echo picked up at the listener side.

If talker echo is reflected twice it can also affect the listener. In this unusual case the listener hears the talker's voice twice: a loud signal first, and then attenuated and much delayed. This is **listener echo**.

These two types of echo are illustrated on Figure 3.4.

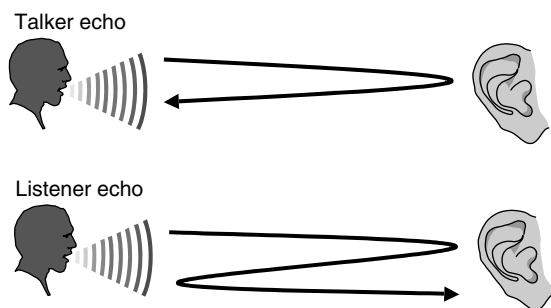


Figure 3.4 Talker and listener echo.

3.3.2 Electric echo

3.3.2.1 What is a hybrid?

The simplest telephone system would look like Figure 3.5. However, to use fewer wires, the phone system was designed to use just two wires. The first 2-wire phones looked like Figure 3.6. Because of parasitic capacities on the line, most microphone signals were dissipated in the talker's loudspeaker (who then tended to speak lower), and almost nothing reached the listener.

The final design arrived at is as shown on Figure 3.7, where Z_{ref} matches the characteristic impedance of the line. Now, the microphone signal is split equally between Z_{ref} and the line, and the speaker hardly hears himself in his own loudspeaker (a small unbalance is kept for him not to have the impression that he is talking in the air). In the ETSI standard Z_{ref} is a 270- Ω resistor connected to a 750- Ω resistor in parallel with a 150-nF capacitor. In France, for instance, Z_{ref} is a 150-nF capacitor in parallel with a 880- Ω resistor, wired to a 210- Ω resistor (complex impedance), but some older phones are also equipped with a real impedance of 600 Ω .

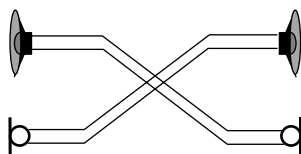


Figure 3.5 Simplest phone network.

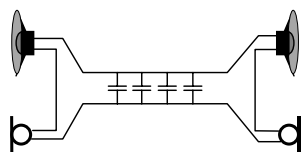


Figure 3.6 Basic phone connection over a single pair.

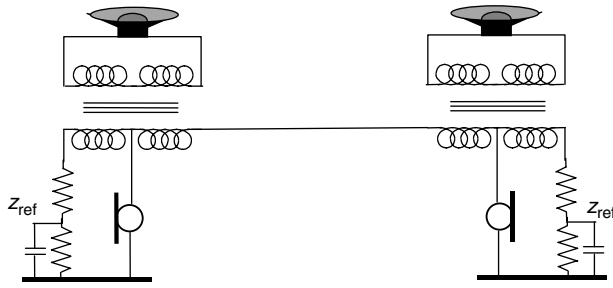


Figure 3.7 Improved design using a hybrid.

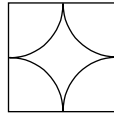


Figure 3.8 Hybrid symbol.

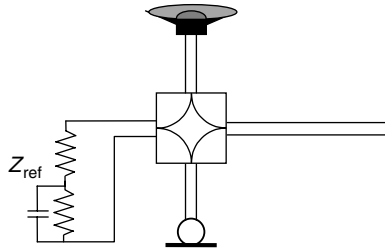


Figure 3.9 Simplified representation of an analog phone.

These values were found to be a good average for a typical line. The actual impedance of a given line will vary according to its length (between 0 km and 9 km, typically), so there is *always* some mismatch.

The common way to symbolize this impedance adaptation device is illustrated in Figure 3.8, where each corner represents 2 wires. It is called a duplexer, or a hybrid. Each half of the circuit of Figure 3.7 can be represented as in Figure 3.9. A hybrid can be integrated easily, a possible circuit is shown in Figure 3.10.

The hybrid is also commonly used in an analog telephone network to allow line signal amplification using the configuration of Figure 3.11.

3.3.2.2 Electric echo

In Figure 3.7 or Figure 3.11, Z_{ref} never matches exactly the characteristic impedance of the 2-wire line, so a portion of the incoming signal is fed back in the outgoing signal. This parasitic signal is the hybrid echo and has all sorts of consequences:

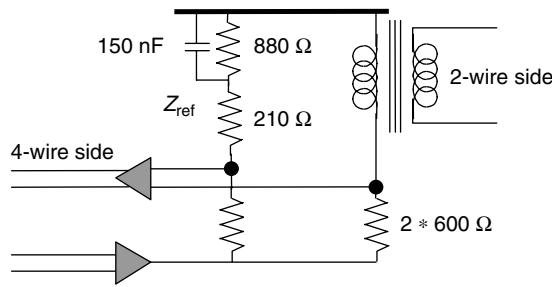


Figure 3.10 Emulating a hybrid with operational amplifiers.

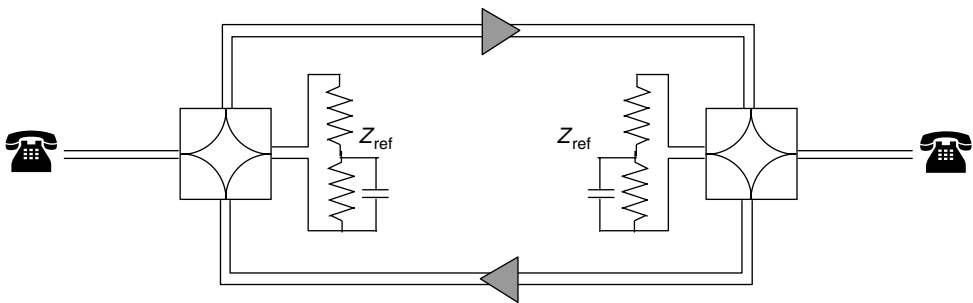


Figure 3.11 Line amplification in the 4-wire path.

- For instance, in Figure 3.11 the signals will loop between the two amplifiers and generate a ‘cathedral effect’ if the one-way delay is about 20 ms. To avoid instability in the network, a loss of 6 dB at least is introduced in the 4-wire path.
- The talker at the other end of the line will hear himself after a round trip delay (talker echo).

In many countries, the transit network is entirely built using 4 wires (any digital link is a virtual 4-wire link). Two- to 4-wire separation occurs at the local switch where the analog phone is connected. Because the echo generated at the switch end comes back to the phone undelayed, it has no effect. On the other hand, the echo generated at the phone end travels back to the other phone through the network (Figure 3.12) and is noticed as soon as the round trip time is above 50 ms (without echo cancellation in the 4-wire path).

ITU Recommendation G.165 provides more details on the handling of hybrid echo.

3.3.3 Acoustic echo

Note: In the following text we will term ‘loudspeaker phone’ an amplified phone without acoustic echo cancellation and ‘hands-free phone’ as amplified phone with acoustic echo cancellation.

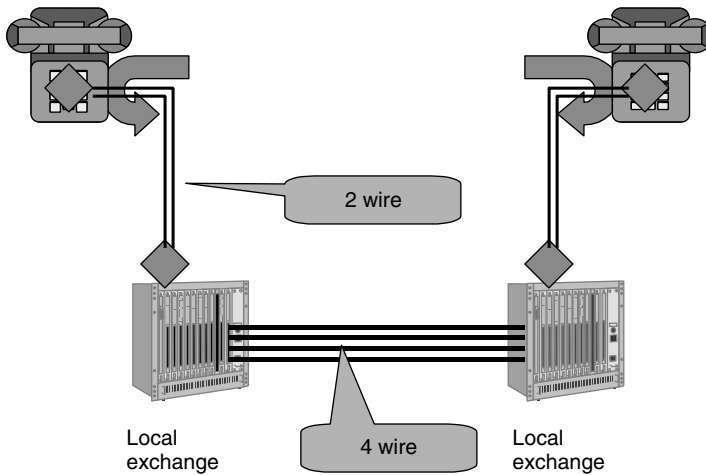


Figure 3.12 Hybrid echo.

Acoustic echo is simply that part of the acoustic signal that is fed back from the loudspeaker of a device to the microphone of that same device. Typically, acoustic echo is a parasitic signal about 10–15 dB (in the case of a loudspeaker phone) below the acoustic signal of the person actually talking into the microphone. Just like hybrid echo, such a level of acoustic echo goes unnoticed if the round trip delay is below 50 ms. After 50 ms the person at the other end of the line gets the impression of talking inside a deep well and then begins to distinctly perceive the echo of his own voice.

An easy way to suppress acoustic echo is to use a headset. However, with appropriate echo-canceling devices it is possible to reduce the power of parasitic echo to about 45 dB below the speaker's signal, even using a loudspeaker phone.

ITU recommendations G.161, G.167, and P.330 focus on acoustic echo and give some values for the typical echo path to use during the testing of echo cancelers:

- for teleconference systems, the reverberation time [time after which the sound energy of an impulse has decayed below 60 dB of the original power] averaged over the transmission bandwidth shall be typically 400 ms. The reverberation time in the highest octave shall be no more than twice this average; the reverberation time in the highest octave shall be no less than half this value. The volume of the typical test room shall be of the order of 90 m³.
- for hands free telephones and videophones, the reverberation time averaged over the transmission bandwidth shall be typically 500 ms; the reverberation time in the highest octave shall be no more than twice this average; the reverberation time in the highest octave shall be no less than half this value. The volume of the typical test room shall be of the order of 50 m³.
- for mobile radio telephones an enclosure simulating the interior of a car can be used.[...] A typical average reverberation time is 60 ms. The volume of the test room shall be 2.5 m³.

Echo cancelers usually do not work as well with acoustic echo as with electric echo, because the acoustic echo path varies much more, which makes it more difficult to dynamically adapt the synthesized echo to the real one. In particular, echo cancelers compliant with ITU recommendations G.165 performance are likely to be insufficient. Even the newer recommendation, G.168, already implemented by most vendors, may not be sufficient in some cases. Both recommendations also provide the ability to stop echo cancellation when detecting the phase reversal tone of high-speed modems.

Typical values for acoustic echo attenuation in current devices are:

- Loudspeaker phones (80% of the market): 10–15 dB.
- Hands-free phones: 35–40 dB.
- Phones with good-quality handsets: 35–40 dB.

3.3.4 How to limit echo

Two types of devices are commonly used to limit echo: echo cancelers and echo suppressors. Electric and acoustic echo reduction is measured in the 4-wire path with the reference points indicated in Figure 3.13.

3.3.4.1 Echo suppressors

Echo suppressors were introduced in the 1970s. The idea is to introduce a large loss in the send path when the distant party is talking. This technique is widely used in low-end hands-free phones, but tends to attenuate the talker when the distant party talks at the same time. It is very noticeable because the background noise that was perceived over the talker's voice by the listener suddenly disappears when he stops speaking or when the listener starts talking. It sometimes creates the impression that the line has been cut, prompting the response: 'Are you still there?'

3.3.4.2 Echo cancelers

The echo canceler functional model is shown in Figure 3.14. An echo canceler is much more complex than an echo suppressor, because it actually builds an estimate of the shape

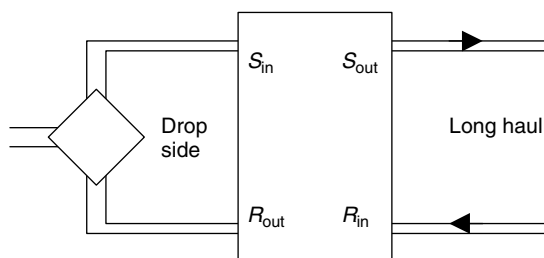


Figure 3.13 Reference points for echo measurement.

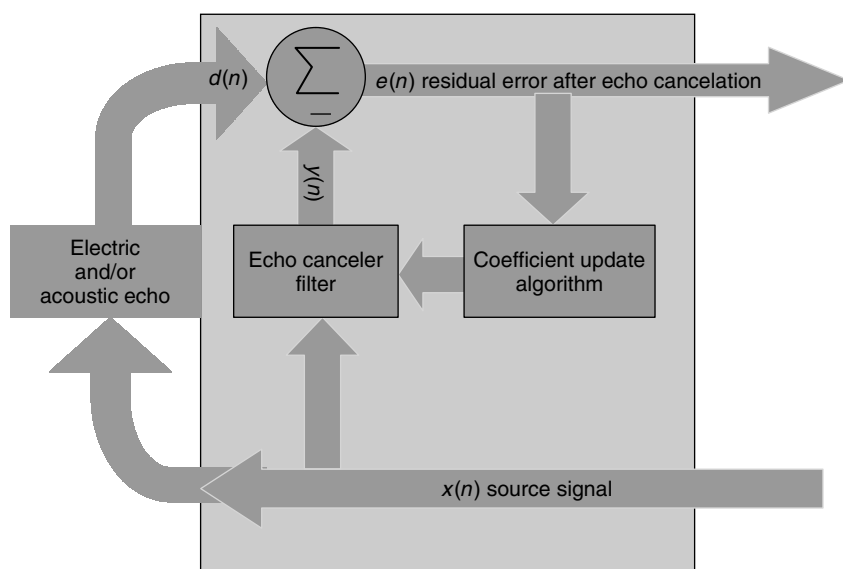


Figure 3.15 Principle behind an echo canceler.

One of the most common algorithms is the recursive least mean squares (LMS) algorithm, which computes the optimal $h(k)$ using a descent algorithm. After each new sample $x(n)$, the $h(k)$ coefficients are updated as follows, where α is the descent algorithm step size parameter:

$$h_p(k) = h_{p-1}(k) + \alpha e(n)x(n - k)$$

A larger step size accelerates convergence while slightly decreasing the quality of echo cancellation. The step size should be smaller than $1/(10 \cdot N \cdot \text{Signal power})$ to ensure stability. Signal power can be approximated by:

$$\frac{1}{M} * \sum_{n=0}^{M-1} x(n)^2$$

3.3.4.3 Usage of echo cancelers

Electric (hybrid) echo cancelers (EECs) are also called line echo cancelers. They are inserted right after the hybrid, located between the 4-wire section of the network (the packetized network in the case of VoIP) and the 2-wire portion (Figure 3.16).

Acoustic echo cancelers are usually implemented in the phone itself.

Many national PSTN networks do not have line echo cancelers due to the relatively small transmission delays. Telephony networks that introduce longer delays can be connected to such PSTNs only through line echo cancelers. For example, in the GSM system, one-way delay is around 100 ms due to:

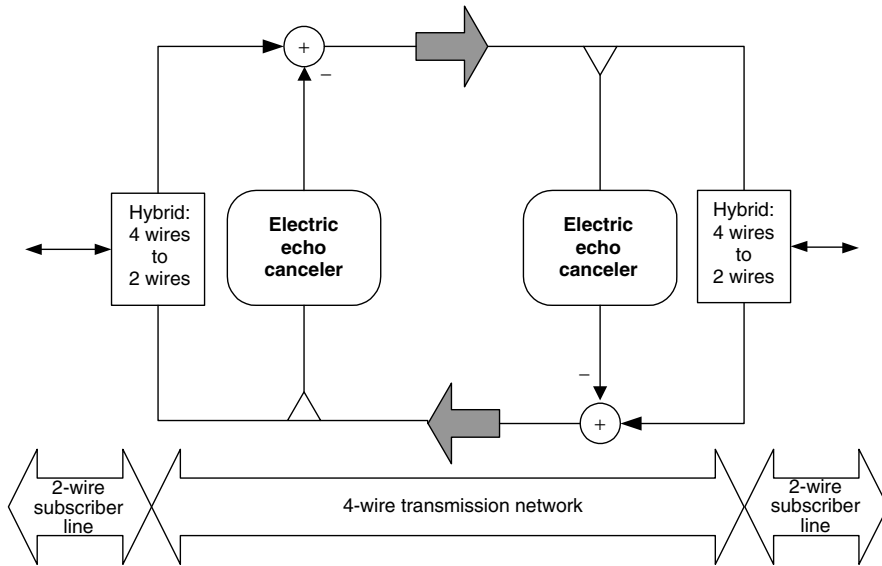


Figure 3.16 Insertion of echo cancelers in a network.

- A frame length of 20 ms.
- A processing delay of about 20 ms (depending on the handset's DSP).
- Interleaving for channel protection.
- Buffering and decoding.

So, an EEC must be included in the mobile switching center (MSC) as shown in Figure 3.17.

The situation is quite similar to that of a VoIP network, where line echo cancellation must be done in the VoIP gateways connected to the PSTN. If line echo cancellation is of insufficient quality, the user on the IP side will hear echo.

VoIP devices can also introduce acoustic echo. The worst examples are PCs with older VoIP software (without acoustic echo cancellation, or AEC). These PCs must be used with headsets in order to reduce echo as much as possible. Note that many headsets are not designed for this (e.g., many headsets have a microphone attached to one of the side speakers, allowing mechanical transmission of speaker vibrations to introduce echo). Some high-end active headsets, as well as dedicated soundboards, now include an AEC module, but the more recent PC VoIP software is now capable of performing the AEC algorithm, making it possible to use standard headsets or even have a hands-free conversation (Figure 3.18).

In a VoIP to PSTN call, if the AEC of the IP phone or the PC is insufficient, echo will be heard at the PSTN end.

The performance of an echo canceler involves many parameters (see G.168 for more details). The most important are echo return loss enhancement, or ERLE (in dB), the amount by which the echo level between the S_{in} and S_{out} port is reduced (see Figure 3.13),

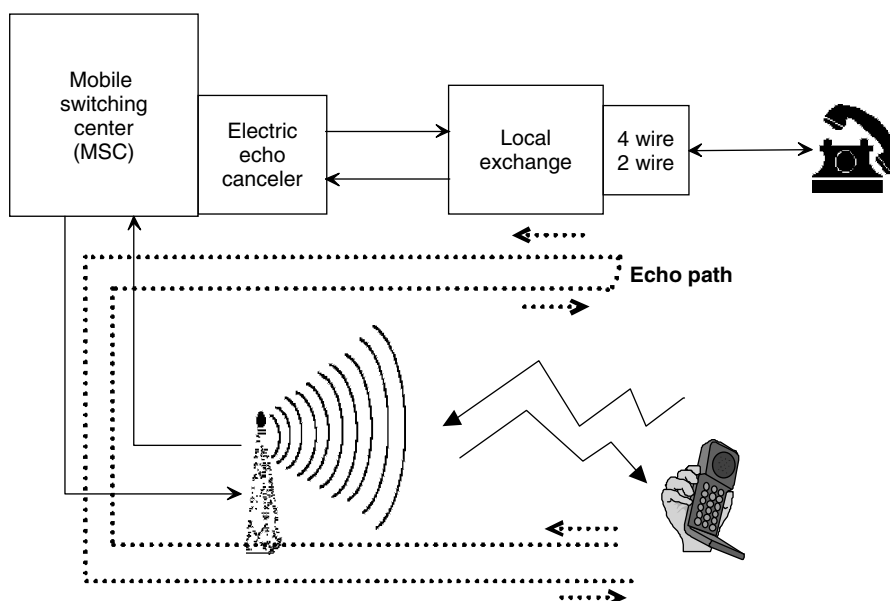


Figure 3.17 EEC required at the interface with the cellular network.

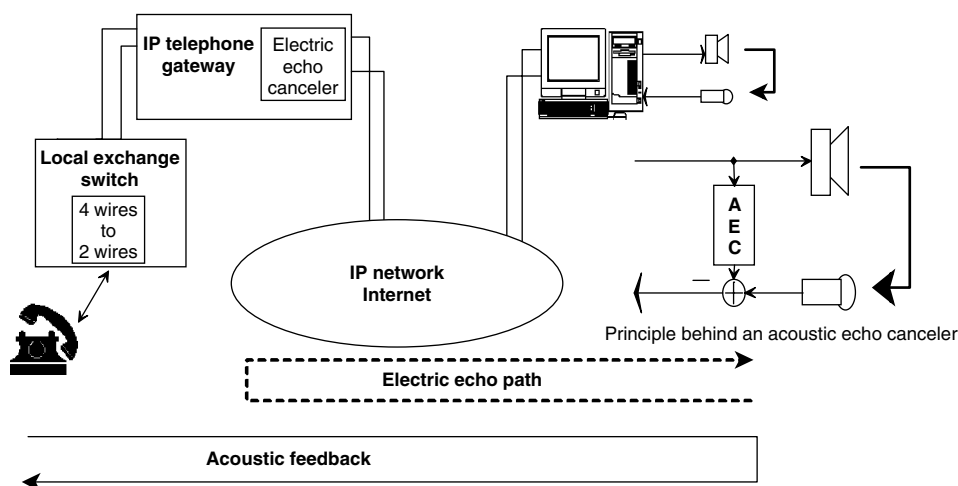


Figure 3.18 Softphones have to provide acoustic echo cancellation.

and the size of the window modeling the impulse response (some echo cancelers are optimized to cancel all echoes coming with a delay of 0 to T_{\max} , some echo cancelers are optimized to model only echoes coming with a delay of T_{\min} to T_{\max}). Other parameters include convergence time and quality of double-talk detection.

3.4 Delay

3.4.1 Influence of the operating system

Most IP phone applications are just regular programs running on top of an operating system, such as Windows. They access sound peripherals through an API (e.g., the Wave API for Windows), and they access the network through the socket API.

As you speak the sound card samples the microphone signals and accumulates samples in a memory buffer (it may also perform some basic compression, such as G.711 encoding). When a buffer is full the sound card tells the operating system, using an interrupt, that it can retrieve the buffer, and stores the next samples in a new buffer, etc.

Interrupts stop the regular activities of the operating system and trigger a very small program called an interrupt handler which in our case may simply store a pointer to the sound buffer for the program that has opened the microphone.

The program itself, in the case of the Wave API, registered a callback function when it opened the microphone to receive the new sample buffers, and the operating system will simply call this function to pass the buffer to our IP phone application.

When the callback function is enacted, it will check that there are enough samples to form a full frame for a compression algorithm, such as G.723.1, and if so put the resulting compressed frame (wrapped with the appropriate RTP information) on the network using the socket API.

The fact that samples from the microphone are sent to the operating system in chunks using an interrupt introduces a small accumulation delay, because most operating systems cannot accommodate too many interrupts per second. For Windows many drivers try not to generate more than one interrupt every 60 ms. This means that on such systems the samples come in chunks of more than 60 ms, independent of the codec used by the program. For instance, a program using G.729 could generate six G.729 frames and a program using G.723.1 could generate two G.723.1 frames for each chunk, but in both cases the delay at this stage is 60 ms, due only to the operating system's maximum interrupt rate.

The self same situation occurs when playing back the samples, resulting in further delays because of socket implementation.

The primary conclusion of this subsection is that the operating system is a major parameter that must be taken into account when trying to reduce end-to-end delays for IP telephony applications. To overcome these limitations most IP telephony gateways and IP phone vendors use real-time operating systems, such as VxWorks (by Wind River Systems) or Nucleus, which are optimized to handle as many interrupts as needed to reduce this accumulation delay.

Another way of bypassing operating system limits is to carry out all the real-time functions (sample acquisition, compression, and RTP) using dedicated hardware and only carry out control functions using the non-real-time operating system. IP telephony board vendors, such as Natural Microsystems, Intel, or Audiocodes, use this type of approach to allow third parties to build low-latency gateways with Unix or Windows on top of their equipment.

3.4.2 The influence of the jitter buffer policy on delay

An IP packet needs some time to get from A to B through a packet network. This delay $t_{AB} = t_{\text{arrival}} - t_{\text{departure}}$ is composed of a fixed part L characteristic of average queuing and propagation delays and a variable part characterizing jitter as caused by the variable queue length in routers and other factors (Figure 3.19).

Terminals use jitter buffer to compensate for jitter effects. Jitter buffer will hold packets in memory until $t_{\text{unbuffer}} - t_{\text{departure}} = L + J$. The time of departure of each packet is known by using the time stamp information provided by RTP. By increasing the value of J , the terminal is able to resynchronize more packets. Packets arriving too late ($t_{\text{arrival}} > t_{\text{unbuffer}}$) are dropped.

Terminals use heuristics to tune J to the best value: if J is too small too many packets will be dropped, if J is too large the additional delay will be unacceptable to the user. These heuristics may take some time to converge because the terminal needs to evaluate jitter in the network (e.g., the terminal can choose to start initially with a very small buffer and progressively increase it until the average percentage of packets arriving too late drops below 1%). For some terminals, configuration of the size of jitter buffer is static, which is not optimal when network conditions are not stable.

Usually endpoints with dynamic jitter buffers use the silence periods of received speech to dynamically adapt the buffer size: silence periods are extended during playback, giving more time to accumulate more packets and increase jitter buffer size, and vice-versa. Most endpoints now perform dynamic jitter buffer adaptation, by increments of 5–10 ms.

A related issue is clock skew, or clock drift. The clocks of the sender and the receiver may drift over time, causing an effect very similar to jitter in the network. Therefore, IP phones and gateways should occasionally compensate for clock drift in the same way they compensate for network jitter.

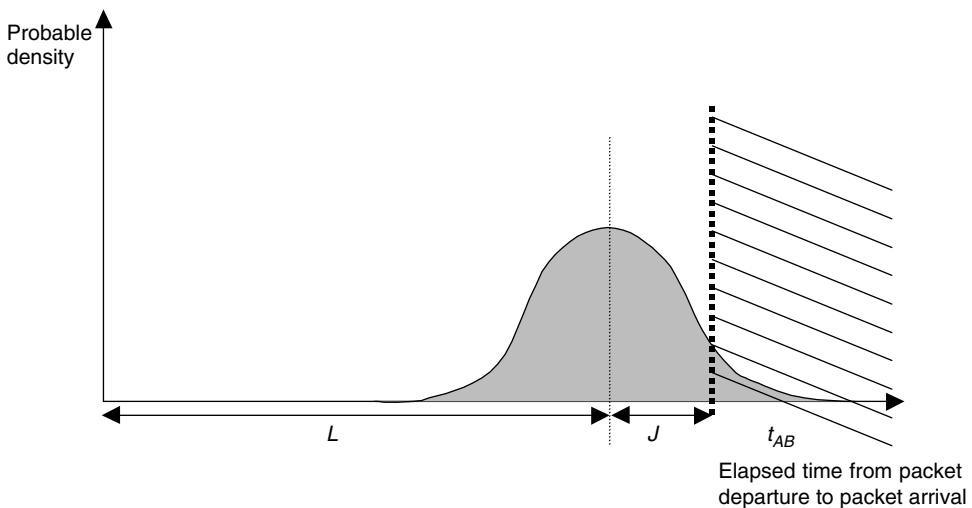


Figure 3.19 Influence of jitter buffer size on packet loss.

3.4.3 The influence of the codec, frame grouping, and redundancy

3.4.3.1 Frame size, number of frames per packet

Most voice coders are frame-oriented; this means that they compress fixed size chunks of linear samples, rather than sample per sample (Figure 3.20). Therefore, the audio data stream needs to be accumulated until it reaches chunk size, before being processed by the coder. This sample accumulation takes time and therefore adds to end-to-end delay. In addition, some coders need to know more samples than those already contained in the frame they will be coding (this is called **look-ahead**).

Therefore, in principle, the codec chosen should have a short frame length in order to reduce delays on the network.

However, many other factors should be taken in consideration. Primarily, coders with larger frame sizes tend to be more efficient, and have better compression rates (the more you know about something the easier it is to model it efficiently). Another factor is that each frame is not transmitted ‘as is’ through the network: a lot of overhead is added by the transport protocols themselves for each packet transmitted through the network. If each compressed voice frame is transmitted in a packet of its own, then this overhead is added for each frame, and for some coders the overhead will be comparable if not greater than the useful data! In order to lower the overhead to an acceptable level, most implementations choose to transmit multiple frames in each packet; this is called ‘bundling’ (Figures 3.21).

If all the frames accumulated in the packet belong to the same audio stream, this will add more accumulation delay. In fact, using a coder with a frame size of f and three frames per packet is absolutely equivalent, in terms of overhead and accumulation

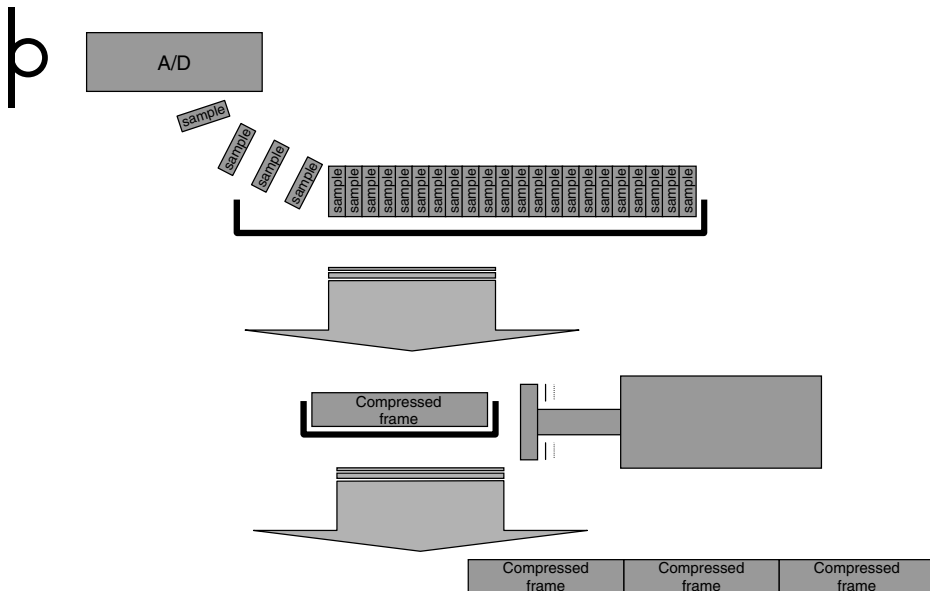


Figure 3.20 Concept of audio codec ‘frames’.

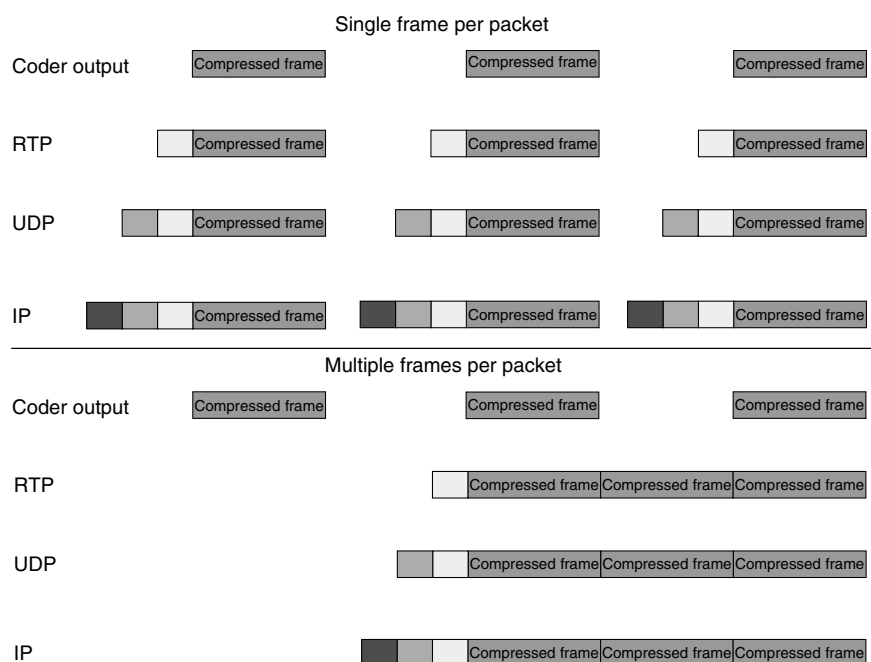


Figure 3.21 Influence of bundling on overhead.

delay, to using a coder with a frame size of $3f$ and one frame per packet. Since a coder with a larger frame size is usually more efficient, the latter solution is likely to be more efficient also.

Note that if the operating system gives access to the audio stream in chunks of size C ms rather than sample per sample (see Section 3.4.1), then samples have already been accumulated and using a coder with a larger frame size f introduces no additional overhead as long as $f < C$.

A much more intelligent way of stacking multiple frames per packet in order to reduce overhead without any impact on delay is to concatenate frames from different audio streams, but with the same network destination, in each packet. This situation occurs frequently between corporate sites or between gateways inside a VoIP network. Unfortunately, the way to do this RTP-multiplexing (or RTP-mux) has not yet been standardized in H.323, SIP, or other VoIP protocols. The recommended practice is to use TCRTP (tunneling-multiplexed compressed RTP), an IETF work-in-progress which combines L2TP (Layer 2 Tunneling Protocol, RFC 2661), multiplexed PPP (RFC 3153), and compressed RTP (RFC 2508, see ch. 4 p. 176).

3.4.3.2 Redundancy, interleaving

Another parameter that needs to be taken into account when assessing the end-to-end delay of an implementation is the redundancy policy. A real network introduces packet

loss, and a terminal may use redundancy to be able to reconstruct lost frames. This can be as simple as copying a frame twice in consecutive packets (this method can be generalized by generating, after each group of N packets, a packet containing the XORed value of the previous N packets, in which case it is called FEC, for forward error correction) or more complex (e.g., interleaving can be used to reduce sensitivity to burst packet loss).

Note that redundancy should be used with care: if packet loss is due to congestion (the most frequent case), redundancy is likely to increase the volume of traffic and as a consequence increase congestion and network packet loss rate. On the other hand, if packet loss was due to an insufficient switching capacity (this is decreasingly likely in recent networks, but may still occur with some firewalls), adding redundancy by stacking multiple redundant frames in each packet will not increase the number of packets per second and will improve the situation.

Redundancy influences end-to-end delay because the receiver needs to adjust its jitter buffer in order to receive all redundant frames before it transfers the frame to the decoder (Figure 3.22). Otherwise, if the first frame got lost, jitter buffer would be unable to wait until it has received the redundant copies, and they would be useless! This can contribute significantly to end-to-end delay, especially if the redundant frames are stored in noncontiguous packets (interleaving) in order to resist correlated packet loss. For this reason this type of redundancy is generally not used for voice, but rather for fax transmissions which are less sensitive to delays.

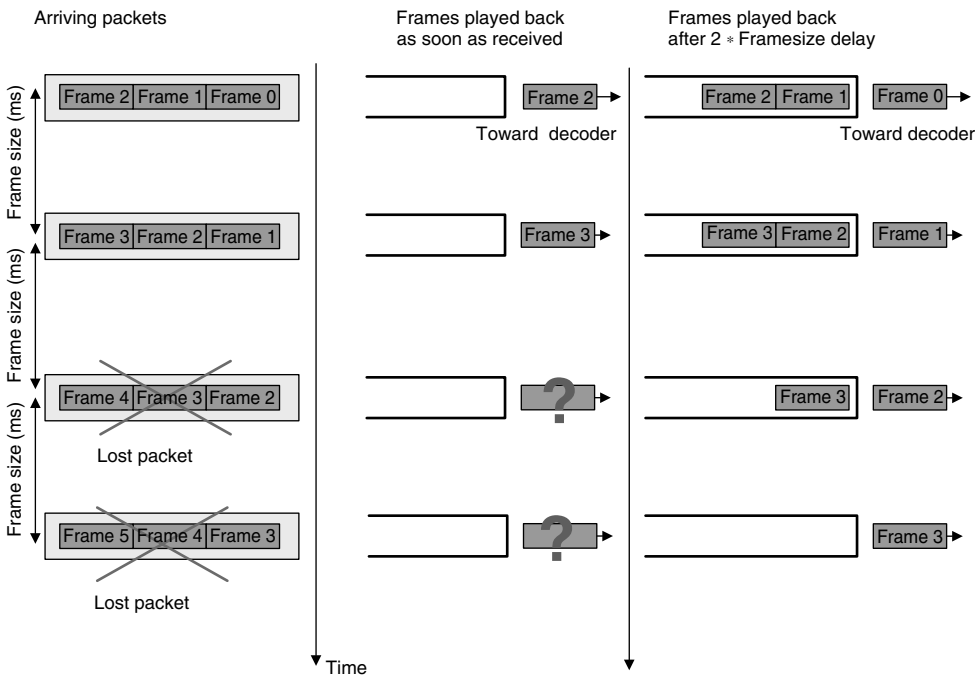


Figure 3.22 FEC or interleaving type of redundancy only works when there is an additional delay at the receiver buffer.

3.4.4 Measuring end-to-end delay

In order to assess the delay performance of IP telephony hardware or software, it is necessary to simulate various network conditions, characterized by such parameters as average transit delay, jitter, and packet loss. Because of the many heuristics used by IP telephony devices to adapt to the network, it is necessary to perform the end-to-end delay measurement after allowing a short convergence time. A simple measurement can be done according to the following method:

- (1) IP telephony devices' IP network interfaces are connected back to back through a network simulator.
- (2) The network simulator is set to the proper settings for the reference network condition considered for the measurement; this includes setting the average end-to-end delay L , the amount of jitter and the jitter statistical profile, the amount of packet loss and the loss profile, and possibly other factors, such as packet desequencing. Good network simulators are available on Linux and FreeBSD (e.g., NS2: www.isi.edu).
- (3) A speech file is fed to the first VoIP device with active talk during the first 15 s (Talk1), then these follows a silence period of 5 s, then active talk again for 30 s (Talk2), then a silence period of 10 s.
- (4) The speech file is recorded at the output of the second VoIP device.
- (5) Only the Talk2 part of the initial file and the recorded file is kept. This gives the endpoint some time to adapt during the silence period if a dynamic jitter buffer algorithm is used.
- (6) The average level of both files is equalized.
- (7) If the amplitude of the input signal is $IN(t)$, and the amplitude of the recorded signal is $OUT(t)$, the value of D maximizing the correlation of $IN(t)$ and $OUT(t + D)$ is the delay introduced by the VoIP network and the tested devices under measurement conditions. The correlation can be done manually with an oscilloscope and a delay line, adjusting the delay until input and output speech samples coincide (similar envelopes and correlation of energy), or with some basic computing on the recorded files (for ISDN gateways the files can be input and recorded directly in G.711 format).

The delay introduced by the sending and receiving devices is $D - L$, since L is the delay that was introduced by the network simulator. With this method it is impossible to know the respective contributions to the delay from the sending and the receiving VoIP devices.

Note that a very crude, but efficient way of quickly evaluating end-to-end delay is to use the Windows sound recorder and clap your hands (Figure 3.23). The typical mouth-to-ear delay for an IP phone over a direct LAN connection is between 45 ms and 90 ms, while VoIP softphones are in the 120 ms (Windows XP Messenger) to 400 ms range (NetMeeting and most low-end VoIP software without optimized drivers).

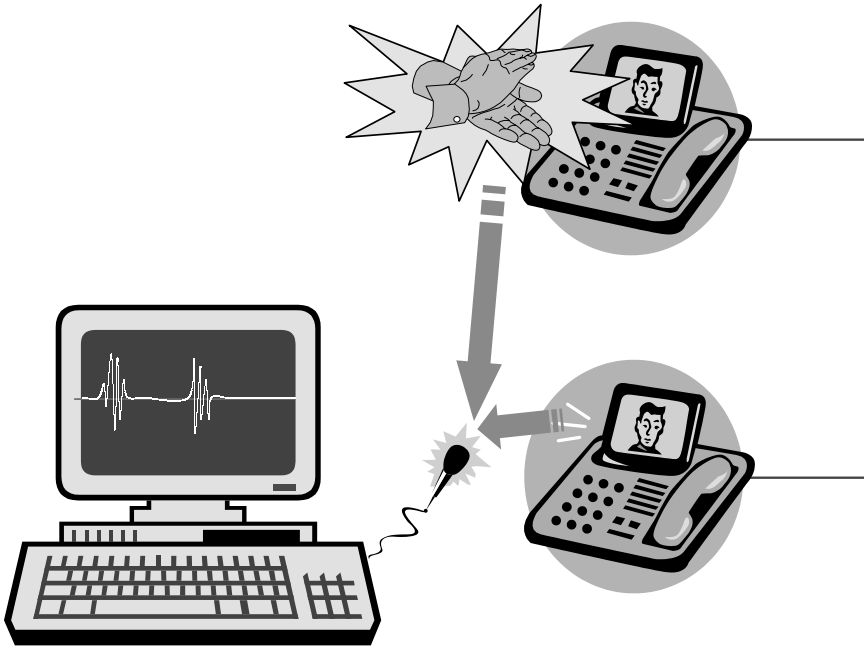


Figure 3.23 Poor man's delay evaluation lab.

3.5 Acceptability of a phone call with echo and delay

3.5.1 The G.131 curve

The degree of annoyance of talker echo depends on the amount of delay between the original signal and the echo, and on the attenuation of the echo signal compared with the original. This attenuation is characterized by the 'talker echo loudness rating' (TELR) as described in G.122 and annex A/G.111. A higher value of TELR represents better echo cancellation. Note that, because of quantization noise on the original signal, it is impossible to achieve a perfect echo cancellation (typically about 55 dB at best).

G.131 provides the minimum requirements on TELR as a function of the mean one-way transmission time T . According to G.131 (Figure 3.24), conditions are generally acceptable when less than 1% of the users complain about an echo problem. The second curve, where 10% of users complain, is an extreme limit that should be allowed only exceptionally.

Figure 3.24 clearly shows that echo becomes more audible as delay increases. This is the reason echo is such a problem in all telephony technologies that introduce high delays. This is the case for most packet voice technologies, for networks that use interleaving for error protection (e.g., cellular phones), and for satellite transmissions in general.

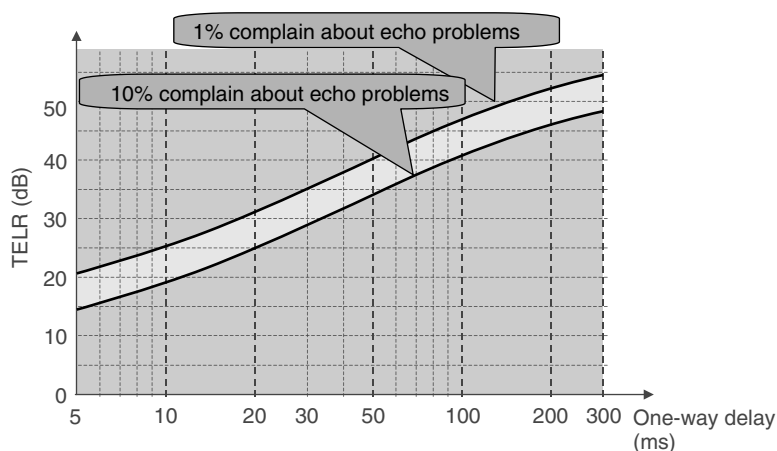


Figure 3.24 G.131 one-way delay versus echo perception.

3.5.2 Evaluation of echo attenuation (TELR)

3.5.2.1 Overview of signal level measurement (dB, dBr, dBm0, etc.)

A discussion of units can be found in G.100 annex A. Here is a short summary:

- **Relative power** is measured in *dB*. A signal of P_1 mW is at level L dB compared with a signal of P_2 mW if $L = 10 \log_{10}(P_1/P_2)$. For relative voltages, currents, or acoustic pressure, the formula uses a multiplicative factor of 20 instead of 10 (power depends on the square of voltage/current or pressure).
- **dBm** refers to a power measurement in dB relative to 1 mW.
- **dBr** is used to measure the level of a reference 1,020-Hz signal at a point compared with the level of that same reference signal at the reference point (the 0-dBr point). For instance, if the entrance of an *2 amplifier (Figure 3.25) is the 0-dBr point, the output is a +3-dBr point. Digital parts of the network are by convention at 0 dBr (unless digital gain or loss is introduced). To determine the dBr level at the analog end of a

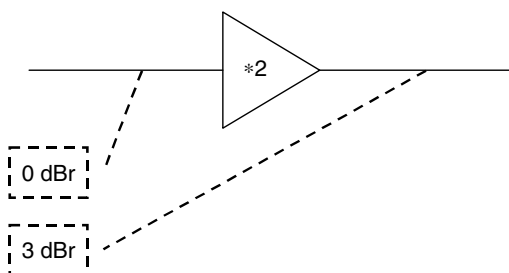


Figure 3.25 dBr levels at the input and output of a *2 amplifier.

coder or decoder, G.101 defines a digital reference sequence (DRS). When decoding the DRS, if the output of the decoder is at R dBm, then it is an R -dBr point.

- **dBm0** is used to measure the absolute level in dBm which a signal would have when passing through the reference point (0-dBr point). For instance, if the power of a signal at the output of the Figure 3.25 amplifier is 10 dBm, then it is a 7-dBm0 signal.

3.5.2.2 TELR for analog and digital termination lines

Recommendation G.131 uses the reference circuit of Figure 3.26 to evaluate talker echo attenuation. The send loudness rating (SLR) and receive loudness rating (RLR) model the acoustic-to-electric efficiency of the emitter and the receiver, respectively (see ITU recommendation P.79). For typical phone sets G.121 states that $SLR_{\text{target}} = 7$ dB, $SLR_{\text{min}} = 2$ dB, $RLR_{\text{target}} = 3$ dB, $RLR_{\text{min}} = 1$ dB. For digital phones, the recommended values are $SLR = 8$ dB and $RLR = 2$ dB.

For an analog phone at the distant side $TELR = SLR + RLR + R + T + L_r$, where R and T stand for additional loss introduced in the analog circuit in order to have a 0-dBr point at the exchange. Most analog phone connections have an $L_r > 17$ dB for an average length of subscriber cable; however, in some networks it can be 14 dB with a standard deviation of 3 dB. In many networks $R + T = 6$ dB.

For a digital phone at the distant side $TELR = SLR + RLR + TCL$, where TCL is terminal coupling loss. IP phones are digital phones. For software phones the values of SLR and RLR can be affected by sound card settings (microphone volume, speaker volume), and properly implemented software should apply digital attenuation to make sure that the resulting SLR and RLR provide the recommended values for the voice level in the VoIP network. Most digital handsets have a TCL of 40–46 dB, although lower end phones may have a TCL as low as 35–40 dB.

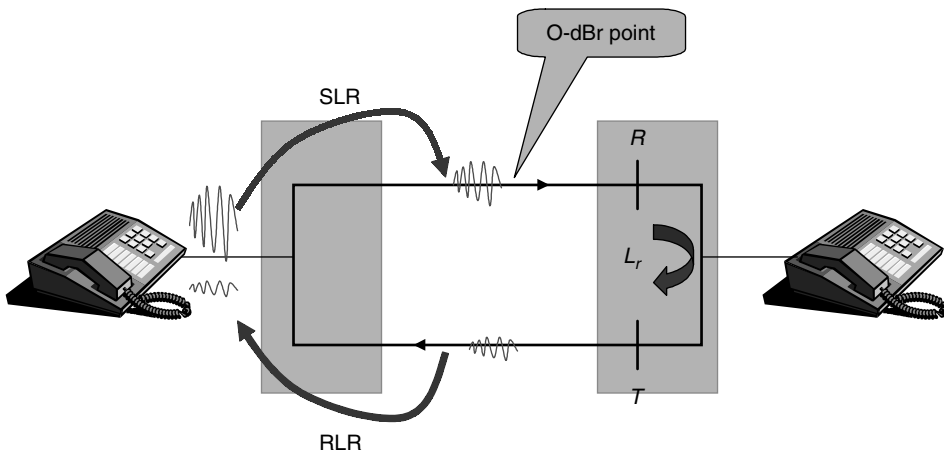


Figure 3.26 Parameters influencing TELR.

When the intrinsic TELR value of the network is too low for the expected network delay, an echo canceler must be added (in the handset for acoustic echo, in the network for line echo) in order to increase the resulting TELR to a value that is acceptable.

3.5.2.2.1 Examples

For $SLR = 7$ dB, $RLR = 3$ dB, $L_r = 14$ dB, $R + T = 6$ we get a TELR of 30 which leads to an acceptable limit for the one-way delay of 18 ms (33 ms in the limiting case). For a 'loud' telephone set with $SLR = 2$ dB, $RLR = 1$ dB, and an L_r of 8 dB we get a TELR of 17 and the limiting case is now 7 ms!

When ringing a digital handset ($TCL = 45$ dB) with the talker's phone at $SLR = 7$ dB, $RLR = 3$ dB we get a TELR of 55 dB and the one-way delay is 'acceptable' up to 400 ms regarding echo perception (but such a one-way delay is already impacting the interactivity of the conversation).

3.5.2.3 VoIP circuits

The IP telephony circuit is subject to the same echo/delay constraints as any other telephony technology. With current IP technology, delays of 200–300 ms for one-way transmission are still common over wide area networks. Other delay factors (encoding delay, jitter buffers, etc.) may add as much as 100 ms. Therefore, all VoIP networks require state-of-the-art echo cancellation, with a TELR value of at least 55 dB. Note that this is close to the highest achievable value for G.711 encoded voice signals, due to the quantization noise that is introduced by G.711. With most echo cancelers, this echo cancellation level can be reduced to about 30 dB under double-talk conditions.

3.5.2.3.1 IP software phone to IP phone or IP software phone

In this case if we assume $SLR + RLR = 10$ dB, then the echo loss of the distant IP phone must be at least $TCL = 45$ dB. On a software phone, this might be implemented in the audio peripherals (soundboard, headset) or by the IP telephony software itself.

3.5.2.3.2 IP phone to a regular phone

3.5.2.3.2.1 IP phone to digital or cellular phone At the ISDN phone end, only acoustic echo is generated since there is no hybrid. Most ISDN phones have a TCL value of 45 dB, so the IP telephony gateway does not need to perform echo cancellation at the ISDN phone end if the connection is digital end to end (this is rarely the case, except in Germany).

Obviously, the IP phone needs to have an echo canceler as well, otherwise the digital phone user will hear echo.

In the early days of VoIP, many gateway demonstrations made phone calls to ISDN or cellular phones. In the case of cellular phones, some vendors even explained that this was the worst case scenario because, after all, you were calling a cellular phone. In fact, this was done on purpose to hide the lack of an echo-canceling algorithm in the IP telephony gateway! The cellular phone itself is a 4-wire device (no electric echo) and includes a

powerful acoustic echo canceler. The cellular phone network interface with the regular phone network is also made via echo cancelers.

3.5.2.3.2.2 IP phone to PSTN user In this case the PSTN phone will generate hybrid echo and acoustic echo. Since propagation time in the PSTN is usually low, many national links may not implement sufficient echo cancellation (if implemented at all). Therefore, the gateway must implement echo cancellation. In some cases there will already be an echo canceler in the PSTN path, which may cause some signal degradation (e.g., voice clipping), but even such degradation is preferable to the risk of not having any echo cancellation.

Canceling electric and acoustic echo is difficult because their characteristics in terms of attenuation and, more importantly, delay are very different. Acoustic echo signal components are typically spread over about 50 ms (office environment), while electric echo signals are typically spread over 13 ms. Echo cancelers are often characterized by the maximum skew between the signals that compose the echo. This signal is a superposition of signals s_i that are a copy of the original signal but attenuated by a factor A_i and delayed by a factor of $D + d_i$. D is introduced by the voice transport network between the echo canceler and the source of echo. If a gateway is implemented in a country the size of France, for instance, D is below 64 ms in 90% of the cases, including call rerouting.

Some echo cancelers optimized for use in corporate equipment only work with $D = 0$ and $0 < d_i < \text{Max skew}$ (e.g., 18 ms). Some network echo cancelers can work with D as large as 500 ms and $0 < d_i < \text{Max skew}$. Only variation of the delay (maximum skew) requires memory in the echo canceler. Since most echo cancelers are implemented as FIR filters on signals that were originally G.711 signals, the memory (therefore, variation of the delay) supported by the echo canceler is sometimes mentioned as ‘taps’ (i.e., a memory cell for one sample, or 0.125 ms). An acoustic echo canceler requires more than 400 ‘taps’ (50 ms), while a line echo canceler requires about 100 ‘taps’ (12.5 ms). Most VoIP gateways have an echo canceler with a memory of at least 32 ms, (many go up to 64 or even 128 ms), and most of them only cancel hybrid echo, which explains why some echo can still be heard sometimes when talking to people with low-quality loudspeaker phones.

Note that it should be possible to disable this echo canceler, either statically (if the gateway is connected to a network already performing echo cancellation) or dynamically (if a modem connection is detected, because modems perform their own echo cancellation as required by recommendation G.168).

3.5.3 Interactivity

In the previous examples the term ‘acceptable’ only considered echo. Interactivity is another factor that must also be considered. Usually, a delay below 150 ms one-way provides good interactivity. A one-way delay between 150 ms and 300 ms provides acceptable interactivity (satellite hop). A one-way delay in excess of 400 ms should be exceptional (in the case of two satellite hops it is about 540 ms) and is the limit after

Table 3.1 ITU interactivity classes

Class	One-way delay (ms)	
1	0–150	Acceptable for most conversations. Only the most interactive tasks will perceive a substantial degradation
2	150–300	Acceptable for communications with low interactivity (communication over satellite link)
3	300–700	Conversation becomes practically half-duplex
4	Over 700	Conversation impossible without some training to half-duplex communications (military communication)

Table 3.2 Communication impairment caused by one-way delay

One-way delay (ms)	Index of communication impairment (%)
200	28
450	35
700	46

which the conversation can be considered half-duplex. ITU recommendation G.114 lists classes of interactivity and quality as a function of delay (Tables 3.1 and 3.2).

When there are large delays on the line, the talker tends to think that the listener has not heard or paid attention. He will repeat what he said and be interrupted by the delayed response of the called party. Both will stop talking . . . and restart simultaneously! With some training it is quite possible to communicate correctly, but the conversation is not natural.

3.5.4 Other requirements

3.5.4.1 Average level, clipping

Gateways and transcoding functions to the PSTN should implement automatic-level control to respect ITU recommendation G.223: ‘The long term average level of an active circuit is expected to be -15 dBm0 including silences. The average level during active speaking periods is expected to be -11 dBm0.’ The methodology for measuring active speech levels can be found in ITU recommendation P.56.

Note that PCM coding is capable of handling a maximum level of $+3.14$ dBm0 in the A law ($+3.17$ dBm0 in the μ law). The gateways should absolutely avoid clipping, since this would adversely disturb the echo cancelers in the network (introduction of nonlinearities). As the average-to-peak power ratio of voice signals is about 18 dB, this imposes an average level not exceeding -15 dB.

Even IP phones and software phones should respect the average levels expected on a transmission line. Since microphone sensitivity can be adjusted on most operating systems, software phones should adjust accordingly to avoid sending too high or too low signals over the VoIP network.

3.5.4.2 Voice activity detection

VAD algorithms are responsible for switching the coder from ‘active speech mode’ to ‘background noise transmission mode’ (this can also be ‘transmit nothing mode’). If they are not implemented properly these algorithms may clip parts of active speech periods: beginning of sentences, first syllables of words, etc.

A general guideline for a good VAD algorithm is to keep the duration of clipped segments below 64 ms and have no more than 0.2% of the active speech clipped. These guidelines are part of ITU recommendation G.116. More detailed information is available in ‘Subjective effects of variable delay and speech loss in dynamically managed systems’, J. Gruber and L. Strawczynski, *IEEE GLOBECOM* ’82, 2, pp. 7.3.1–7.3.5.

3.5.5 Example of a speech quality prediction tool: the E-model

The E-model was originally developed in ETSI for the needs of network planning and later adopted by the ITU as recommendation G.107. It allows the subjective quality of a conversation as perceived by the user to be evaluated. The E-model appraises each degradation factor on perceived voice quality by a value called an ‘impairment factor’. Impairment factors are then processed by the E-model which outputs a rating R between 0 and 100. The R value can be mapped to a mean opinion score, conversational quality evaluation (MOS_{CQE}) value between 1 and 5, or to percent good or better (%GoB), or to percent poor or worse (%PoW) values using tables. An R value of 50 is very bad, while an R value of 90 is very good.

The E-model takes into account parameters that are not considered in the G.131 curve (Figure 3.24), such as the quality of the voice coder (degradation factor I_e) and frame loss (degradation factor B_{p1}). Most voice coders have been rated for their impairment factor without frame loss, and consequently the E-model (available as commercial software from various vendors) can readily be used to evaluate perceived voice quality through an IP telephony network with no packet loss and low jitter. This work was published by the T1A1.7 committee in January 1999.

Impairment factor parameters are evaluated from real subjective tests to calibrate the model (e.g., see G.113). Therefore, the usability of the E-model for a particular technology depends on how much calibration has been done previously on this technology. The E-model is only useful if it is used correctly. An impairment factor for a coder measured under specific loss profiles is not valid for other loss profiles (e.g., if there is correlated loss). IP telephony introduces many new types of perturbations that do not exist on traditional networks, such as the delay variation that may be introduced by endpoints trying to dynamically adjust the size of jitter buffers, voice clipping introduced by VAD algorithms, or correlated loss introduced by frame grouping. Some R&D labs that specialize

in voice quality and network planning have released new versions of the E-model with specific support for IP telephony degradations. This should lead to an update of the ITU specification in 2005 (currently known as P.VTQ).

One of the interesting aspects of the E-model is that it also takes into account psychological parameters that do not influence absolute voice quality, but the *perception* of the user. For instance, the ‘expectation’ impairment factor takes into account the fact that most users expect to have degraded voice quality when using a cellular phone, and therefore will be more indulgent ... and complain less, for the same level of quality, than if they had been using a normal phone and vice versa: if a cellular phone achieves similar voice quality to a normal fixed phone, many users will actually find it better than the normal phone. IP phone manufacturers may have to find a recognizable design if they want to benefit from the ‘VoIP expectation factor’!

3.6 Conclusion

In many ways, IP phone networks and mobile phone networks (such as GSM) face similar constraints regarding voice quality. The main issue in both systems is to correctly control echo, minimize the degradations introduced by packet loss, and preserve good interactivity of speech.

In early VoIP systems, all three factors were a problem and, unfortunately, this brought about a bad perception of VoIP which still persists today:

- Packet loss and delay: in 1998, the Internet was still perceived as a gadget by most incumbent carriers; as a result there was typically too little capacity installed for IP traffic. In addition, IP networks were frequently built on frame relay networks, which introduced long delays (frame relay switches often have large buffers). This situation was reversed after the Internet bubble. The backbones carrying the IP traffic today use state-of-the-art technologies, such as MPLS, and are capable of handling large volumes of traffic with minimal delay. As IP traffic dominates any other kind of data, fewer and fewer encapsulation layers are used to transport IP packets. The frame relay transport layer has been completely abandoned in core backbones, unable to offer the required capacity and performance. Even the intermediary ATM transport layer is becoming a problem at the speed at which many backbones operate today. Many backbones now carry IP packets directly on top of a layer 2 technology, such as SDH. Finally, most VoIP gateways and IP phones now implement sophisticated packet loss concealment methods which can efficiently mask up to two consecutive frames lost, making the occasional lost packet less perceptible.
- Echo control: early VoIP gateway implementations often had low-quality echo canceler implementations. With the consolidation of the telecom market, most gateway manufacturers are either previous DCME equipment vendors (compression equipment for submarine cables) with extensive know-how in signal processing and echo cancelation, or companies using the reference algorithms proposed by their DSP vendors (during the telecom bubble, most DSP vendors either developed internally or acquired companies

that had a lot of know-how in signal-processing algorithms). As a result the quality of echo cancelation in VoIP gateways and IP phones is now much better, reaching the level of echo attenuation required for long-delay networks.

In early VoIP systems, mostly targeted for the prepaid market, there was a lot of attention paid to proprietary redundancy schemes or supposedly high-performance proprietary coders. This trend has now come to an end for many reasons. Many redundancy schemes were promoted for marketing reasons, but didn't perform as advertised in real networks. More fundamentally, as VoIP developed beyond the prepaid market, into business trunking (connection of corporate PBXs to VoIP backbones), residential telephony or IP Centrex, the key requirement became interoperability. VoIP networks using proprietary schemes were unable to evolve and offer new applications, and many disappeared.

Whether or not VoIP networks need to pay attention once more to redundancy algorithms and high-performance low-bitrate coders is debatable. Most wired IP networks now support differentiated quality of service, which means VoIP transmissions enjoy very low packet loss even without any redundancy in the RTP stream. For wireless networks (e.g., UMTS), the trend is to provide various levels of error protection directly at the physical level, dynamically for each type of stream (for IP streams, desired transport layer behavior may be signaled via DiffServ marks).

Future work on voice quality on IP networks will focus on providing better than toll quality on wired lines (such as wide-band voice or stereo/spatialized voice) and improving the quality of voice over IP on wireless networks through tight integration with the QoS mechanisms of the physical layers of UMTS or WiFi networks.

3.7 Standards

[G.100, ITU]	Definitions used in recommendations on general characteristics of international telephone connections and circuits
[G.107, ITU]	The E-model, a computational model for use in transmission planning.
[G.113, ITU]	Transmission impairments.
[G.114, ITU]	One-way transmission time.
[G.116, ITU]	Transmission performance objectives applicable to end-to-end international transmissions
[G.122, ITU]	Influence of national systems on stability and talker echo in international connections.
[G.131, ITU]	Control of talker echo.

[G.122, ITU]	Influence of national systems on stability and talker echo in international connections
[G.111, ITU]	Loudness ratings in an international connection.
[G.168, ITU]	Digital network echo cancelers.
[G.167, ITU]	Acoustic echo controllers.
[G.165, ITU]	Echo cancelers.
[G.174, ITU]	Transmission performance objectives for terrestrial digital wireless systems using portable terminals to access the PSTN
[P.310, ITU]	Transmission characteristics for telephone band (300–3,400 Hz) digital telephones.
[P.79, ITU]	Calculation of loudness ratings for telephone sets.
[P.56, ITU]	Objective measurement of active speech levels.
[P.11, ITU]	Effect of transmission impairment.
[G.175, ITU]	Transmission planning for private/public network interconnection of voice traffic
[G.173, ITU]	Transmission planning aspects of the speech service in digital public land mobile networks
[G.111, ITU]	Loudness ratings in an international connection
[IEEE]	Subjective effects of variable delay and speech loss in dynamically managed systems”, J. Gruber and L. Strawczynski, IEEE GLOBECOM '82, Vol 2: F.7.3.1-F.7.3.5
[ETSI, TR 101329 V 1.2.5]	Telecommunications and Internet Protocol harmonization over networks (TIPHON) : General Aspects of Quality Of Service (QoS)
[Technical Report 56, T1A1.7]	Performance guidelines for voiceband services over hybrid IP/PSTN connections.
