IP Telephony and Some Factors that Influence Speech Quality

Hans W. Gierlich
Vice President
HEAD acoustics GmbH

Introduction

This paper examines speech quality and Internet protocol (IP) telephony. Voice over IP requires specific treatment of the signals transmitted in order to avoid the problems reported by a study conducted by KPN Netherlands and reported at the ETSI-STQ workshop "Quality issues for IP-telephony" June 1999: A pilot study was organized in 1998/1999. The requirement was that the network quality should be better than global system for mobile communications (GSM), i.e. the subjective quality rating expressed in MOS (mean opinion score) should be >3.2. There were about 5,000 customers in the study. Problems were reported due to poor speech quality, echo, clipping, inaudibility and soft speech. Up to Forty-five percent of the users were very dissatisfied with the speech quality and 33 percent were not satisfied with the service. Replacement of the ordinary telephone is not possible using the technology that has been deployed there. It is important to note that all of the complaints were related to speech quality. Call-setup times and availability were of minor importance.

Parameters influencing the Speech Quality

Speech quality is a fairly complicated issue: Speech quality includes parameters determining the conversational situation, the listening situation and the talking situation. Several speech-quality parameters are shown in Figure 1.

Figure 1: Speech-Quality Parameters (slide 2)

Besides the parameters naturalness and speech-(sound) quality the listening-effort influences speech quality. Speech intelligibility has become more of an issue due to coding, speech switching, and various kinds of signal processing. There are noisy environments in which people seldom used the telephone in the past but where it is quite common to use telephones today, telephoning on the street, at the airport, in cars or in railway stations may be examples where people use mobile phones. The environmental situation influences quite significantly the perceived speech quality. In addition there is the conversational aspect of speech in which people can be talked with and interrupted. How people listen and interact is very important, conversational parameters like double talk capability and conversational effort subjectively determine the speech quality in this situation. So what would a customer accept or not accept in terms of quality? If on a mobile connection, for example, would that customer accept more degradation of a signal than if it were a fixed network? If it were an IP network, would the customer accept more signal degradation than with a traditional PSTN? All of this influences the speech quality perceived by the user.
Also it should be noted that speech quality perception may be language dependent. Chinese speech, for instance, is different than American speech, and signal processing may have a different impact.

**Signal Processing in IP-Configurations**

Look at the typical processing in an IP terminal or an IP gateway (see Figure 2) from a speech quality point of view. Starting from the network side there will be the packetizing and buffering. There is the coding of the sent and received speech. There will be signal processing to have the signals separated in double-talk situations, and there will be equalization and voice-activated speech switching. The future is expected to have mostly hands-free type of telephone applications, which will require some control of the acoustic echo coming from the microphone to the loudspeakers. There will be a very sophisticated acoustic echo cancellation in combination with voice activated gain switching. There will be many things that are neither linear nor time invariant.

*Figure 2: IP–Terminal Signal Processing* (slide 3)

**Impact on Speech Quality**

A problem specifically to IP networks is that of packet loss. Packet loss may be unavoidable, it depends on the network load which is impossible to predict. Packet loss results in cutting off a speech segment and is simply a missing block. The effect is different to the well know front-end clipping. Clipping may occur at any time, and the length, duration, and time distribution is fairly unknown.

Another effect may occur when people want cut off pauses in the speech signal, typically in order to reduce the transmission bandwidth needed. Even for undistorted signals this type of signal processing may lead to speech clipping. The problem even gets worse in the presence of background noise. In this case besides the problem of speech detection in the presence of background noise the insertion of a suitable comfort noise into segments where the signal is cut off is of relevance for the speech quality. Furthermore any clipping may interact with the speech coder used in the individual connection and may result in further impairments. From the subjective point of view neither speech clipping nor any background noise variation nor any impairments resulting from improper coding should be noticeable (see Figure 3).

*Figure 3: Packet Loss and Coding* (slide 4)

The most significant parameters describing speech quality are: delay and echo, clipping, the quality of the background-noise transmission (how this background-noise signal is transmitted is quite important for the perceived quality of the connection). The double-talk performance is also important as are echo disturbances under single- and double-talk situations. Loudness and noise are prominent telephony parameters as well.

When assessing the numbers relevant to speech quality, one must first look for the echo and switching, besides the well known traditional parameters like Loudness Ratings and frequency responses the most disturbing parameter in single talk. There is no transmission without delay, and ITU-T Recommendation G.131, as shown in Figure 4, shows the required echo loss in terms of delay in single talk conditions.

*Figure 4: Echo and Delay* (slide 11, right side)
General requirements for switching in single talk conditions can be found e.g. in ITU-T Recommendation P. 340.

**Requirements for Echo and Switching during Double Talk**

What happens in the critical double-talk situation in complex systems? In double talk, the echo-loss requirement can be relaxed, but it is not a single number. When asking about disturbance caused by echoes in a subjective test, a rating of MOS (mean opinion score) > 4.0 is basically the best that can be obtained in a single-talk situation, and while it would be nice to have the same rating in a double-talk situation, the echo-loss requirement can be relaxed (see Figure 5). How much the results are relaxed has something to do with the expectation of the user. If the user believes the connection is of high quality, then a high echo loss is required. The same is valid for the switching characteristics. The switching parameters during double talk are extremely important when judged subjectively. If there is switching loss inserted between single talk and double talk the amount of attenuation is important. High switching loss between single-talk and double-talk situations decreases the rating. E.g. a switching loss of more than 20 dB results in a rating of MOS (mean opinion score) < 1.5 which is very poor (MOS =1 would be totally unacceptable).

*Figure 5: Echo During Double Talk* (slide 12, upper right)

**Standards and Recommendations**

Standards for test signals and procedures have been recommended by the International Telecommunications Union (ITU). Most important are the P.50, P.501 and P.59 which describe test signals useful for objective determination of speech quality. P.340, P.502 and P.861 describe various objective test methods. P.861, "PSQM" for example, describes how, from speech, one number can be derived that determines the speech quality for one way transmission from network termination point (NTP) to NTP. The test procedures described in P.502 and P.340 allow a much more detailed investigation of the various objective parameters. The description of test setups can be found in P.581 and P.64.

In ETSI the project TIPHON is concerned with IP telephony. The TR 101329 (currently in revision) describes test procedures for objective speech quality assessment. The new version will outline separate standards on measurement and quality of service (QoS) parameters. Another ETSI standard, EG 201 377-1, describes test methods for NTP-to-NTP connections but does not include the terminal. The terminal however determines the speech quality to a great extent and it should be recognized, that the terminal and network can no longer be separated, especially in IP scenarios. Such it is advisable to test complete configurations, including terminal and network.

**Example Measurements**

Measurement results were done on two IP configurations. In the first configuration, one side contained the analog inputs and the other a handset (see Figure 6). There was background noise, as usual. There was no packet loss. The voice-activity detection (VAD) was active but set at a very low threshold.

*Figure 6: Configuration One* (slide 14)
The other configuration was a back-to-back connection of two personal computers (PCs) with software solutions. There was electrical access, and a headset was used, as shown in Figure 7. No traffic was on the network.

**Figure 7: Configuration Two (slide 15)**

Some simple measurements using speech like signals led to the following results for delay and echo: For configuration one the overall delay of that connection was 70 milliseconds, the measured echo loss was > 40 dB so it was fairly good. For configuration the delay varied between 480 ms and 540 ms, the measured echo loss was only 21 dB. Both delay and echo loss is not acceptable for a good speech quality, the delay is too high, the echo loss is completely insufficient (the required echo loss in such a connection should be >56 dB).

More detailed investigation of additional parameters led to the following results (examples):

**- Background Noise Transmission**

For the evaluation of background noise transmission a noise like signal with constantly increasing level was used.

For configuration one, the background-noise signal except for low levels is transmitted with no artifacts. Low background noise levels are cut off.

For configuration two the background noise is transmitted fairly incomplete. Independent of the background noise level, the signal is switched on and off which results in high background noise bursts.

**- Level dependent Transmission of Speech Signals**

For this test a voiced sound (speech like) of speech with constantly increasing level was used. In sending direction the signal was transmitted with no problems by both configurations. In receiving direction however the behavior was quite different. Whereas in configuration the signal is transmitted with nearly no artifacts (except switching off for low signal levels), configuration 2 shows a very strong companding of the speech signal. For nearly all input signal levels the output signal level is kept almost at the same level (see Figure 8).

**Figure 8: Transmission of background noise for configuration 2 in receiving direction**

*Lower graph: input signal level vs. time*
- **Double Talk Performance**

Echo during double talk was evaluated using a voice-like test signal consisting of voiced sounds with orthogonal distributed spectra, fed in simultaneously in sending and receiving direction. The echo analysis was made by spectral extraction of the echo components from the double talk sequence.

Using this technology for configuration one it can be shown that the echo loss during double talk is still sufficient (>40 dB). For configuration 2 the sending direction was attenuated by more than 20 dB and such the echo loss is in the range of 40 dB. Thus double talk is impossible due to the high attenuation of the sending direction.

For evaluating the switching parameters during double talk specific speech like test signals were used. They consist of Composite Source Signals (see ITU-T Rec. P.501), overlapping in time with constantly increasing and decreasing signal levels, fed in simultaneously in sending and receiving direction. This signal can be used to evaluate switching during double talk.

In this double-talk test for configuration one the signal transmission was nearly complete. For signal levels in the range from -4,7 dBPa to -20 dBPa the sending direction was transmitted nearly completely. In receiving the signal was transmitted with no clipping in the level range of -8 dBm0 to -28 dBm0. For lower signal levels front-end clipping occurred directly after double talk periods. The measured switch over times however were only in the range of 80 to 150 ms which (for the low signal levels) is sufficient to guarantee a good speech performance during double talk.

For configuration 2 the transmission during double talk sometimes is possible with high echo (the echo loss is still only 21 dB!) and sometimes the sending signal is attenuated by 20 dB. Both, echo and switching are unacceptable for a good double talk performance.

**Summary**

Speech quality is influenced by many things, including the condition and load of the network. This is IP-specific in that it is not normal in standard networks. Interaction of terminal and network components becomes more and more important. The environmental conditions where the terminals are used, and especially the types of terminal designs, is very important in perceived speech quality. Test methods and standards are available for various parameters, but investigations are still necessary to determine overall quality.