Conversational Speech Quality - The Dominating Parameters in VoIP Systems

H. W. Gierlich; F. Kettler
HEAD acoustics GmbH
Ebertstraße 30 a, 52134 Herzogenrath, Germany

Abstract

Speech quality, especially in voice over IP systems is influenced by many parameters:

Delay has significant impact on the conversational quality mainly for two reasons: On the one hand the conversation between both subscribers and their interaction is more difficult. On the other hand the delay has a strong influence on the echo perception, longer transmission delay leads to an increased sensitivity in the users echo perception. It is important to evaluate the echo performance of a system in all conversational situations since typically additional artifacts like voice switching and background noise modulation may occur due to echo cancellation processes.

Voice switching may be introduced at many stages in a VoIP scenario. Voice switching may lead to speech quality degradation perceived as missing syllables or missing words. Since again the effect is subjectively perceived different in single talk situations as compared to double talk situations both situations have to be taken into account.

The transmission of background noise is a very critical parameter for the naturalness of a conversation. Due to clipping caused e.g. by VAD or the use of background noise reduction algorithms, background noise modulation - in combination with comfort noise injection - may degrade the perceived quality significantly. Packet loss may lead to speech quality degradation during single talk and double talk periods.

The article describes objective methods which allow the assessment of various speech quality parameters relevant for the conversational quality.

A. INTRODUCTION

In modern IP systems the communication using speech can no longer be regarded as it used to be in traditional PSTN networks. The differentiation between terminals and network is no longer possible. Typically the same signal processing algorithms are used in networks and in terminals. Basically three different configurations have to be taken into account:

IP gateway to IP gateway , IP terminal to PSTN terminal and IP terminal to IP terminal including the acoustical interfaces like hands-free phones, headsets or handset (see figure 1).

Since the interaction of the signal processing between terminal and network highly influences the speech quality, typically a complete configuration has to be considered for all testing including network and traffic load simulations.

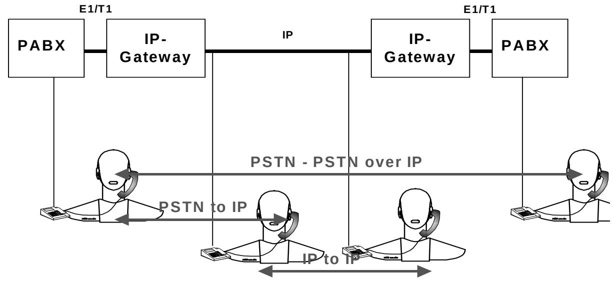

Fig. 1: The typical scenarios in VoIP networks

B. SIGNAL PROCESSING IN IP CONFIGURATIONS

Figure 2 introduces a typical block diagram for the signal processing from a speech transmission quality point of view.

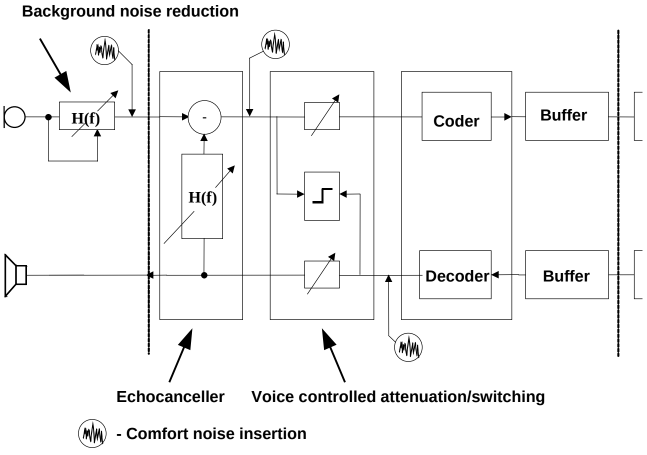
Background noise reduction
H(f)
-
H(f)
Coder
Buffer
Decoder
Buffer
Echocanceller
Voice controlled attenuation/switching
- Comfort noise insertion

Fig. 2: Block diagram of typical components in voice over IP systems

The acoustical access is realized by acoustical components such as hands-free units, headsets or handset terminals. In any case sufficient echo control has to be provided in order to guarantee a sufficient echo loss even under worst case conditions. Due the influence of delay on echo perception (see[1]) these worst case conditions have to take into account the maximum expected delay due to the propagation delay, the delay variation (‘jitter’) in the network, packetization delay and buffer size in the equipment. The echo control can either be achieved by good acoustical terminal design or by using a speech echo canceller or any sort of voice activated switching (echo suppressor) which is shown in the block diagram. Since in many cases the amount of echo cancellation is not a sufficient, speech echo cancellers typically are backed up by voice controlled amplification/attenuation systems in sending and receiving direction. Occasionally cancellation and switching is not realized in full bands but subband echo cancellers and attenuations are used. Sometimes additional signal processing components such as companding devices (AGC) etc. are introduced in order to enhance speech quality under special conditions e.g. noisy environments or a wide range of speech signal levels in the network. Typically voice activity detectors (VAD) are used in order to measure the voice activity and avoid the transmission of speech packets in case no voice activity is detected. In order to hide those silence intervals, comfort noise is inserted (typically on the far end side of a connection to save transmission bandwidth). In any case speech is coded using various sorts of a speech coding algorithms such as G.711, G.723, G.729. Buffers of variable size are used in order to guarantee the transmission of the speech with a minimum of packet loss under typical network and traffic conditions.

C. THE MOST DOMINATING PARAMETERS DETERMINING THE SPEECH QUALITY IN VOIP SYSTEMS

One of the dominating parameters influencing speech quality is delay. Delay may strongly impact the conversational quality but delay as well influences the delectability and annoyance caused by echo. Echo may occur in single talk as well as in double talk situations. In general the echo attenuation required to avoid customer complains is a function of delay and as soon as one way delay exceeds 250 ms, the echo attenuation has to be at least 46 dB. More information can be found in [1]. Due to various kinds of signal processing echo attenuation depends on robustness and adaptation speed of echo cancellers, the reliability of double talk detection, the sensitivity against background noise and other kinds of disturbances. Consequently the requirements on echo attenuation have to be checked under all these conditions. Specifically the double talk situation has to be taken into account. In order to assess this situation subjectively and find limits for the double talk situation auditory tests were conducted recently [2]. Specific double talk tests developed for this specific scenario were used. The test subjects had to rate the annoyance caused by echoes during double talk using MOS (Mean Opinion Scores) MOS 1 corresponds to a highly annoying echo, MOS 5 corresponds to "echo not perceptible". In Fig. 3 the difference in echo loss -always compared to the single talk situation- is shown in relation to the MOS rating.

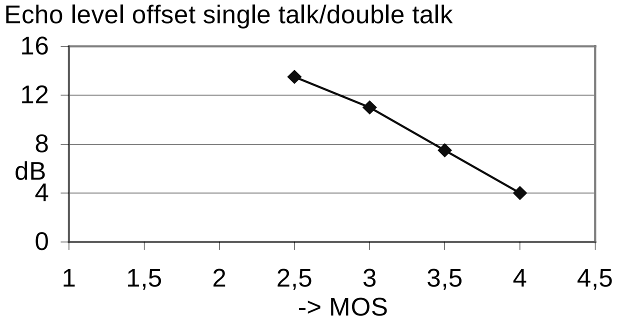

Fig. 3: Differences in echo loss requirement between the single and double talk situation as a function of perceived echo annoyance during single talk

From this results it is obvious that the requirements for echo attenuation during double talk are high and depend on the quality achieved in single talk situation. This means that in case the echo attenuation in single talk situations is poor the requirement for double talk may be low as well. From Figure 3 the echo loss requirement during double talk depending on the annoyance of echo perceived subjectively under single talk conditions can be derived .

Switching itself is a critical parameter under both single and double talk conditions. Basic requirements for switching in the single talk mode are known already since many years (e.g. ITU-T Recommendation P.340 [3]). The annoyance caused by switching, specifically front-end clipping was widely investigated by Sotscheck [4]. The most important result of these investigations is that front-end clipping of more than 15 ms should be avoided in any case. Frontend clipping is certainly not the only relevant parameter for VoIP scenarios since packet loss causes more statistically distributed temporal clipping. Without any kind of error correction (packet loss concealment), the speech gaps resulting from packet loss cause a high degradation of speech sound quality and may even reduce the intelligibility of speech significantly.

More critical however is switching during double talk. Since during double talk typically echo cancellation is less efficient again switching is introduced by additional nonlinear processes in conjunction with the echo canceller function (e.g. loss insertion to guarantee the necessary overall echo attenuation). If a loudness variation between the single and the double talk situation is more than 3 dB, the speech quality is degraded. Figure 4 gives the relationship between the annoyance caused by loudness variations as a function of attenuation range. Again specific double talk tests have been used for the tests. MOS 1 corresponds to highly annoying switching whereas MOS 5 corresponds to switching not perceptible. For more information again see [2].

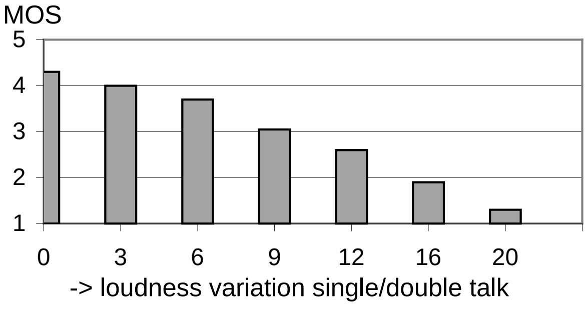

Fig. 4: Annoyance caused by loudness variations between single talk and double talk

When regarding the speech sound quality the speech quality may be highly influenced by packet loss and jitter. Packet loss in combination with various codecs may degrade the speech quality significantly. In order to evaluate this degradations methodologies based on the human perceptions of speech quality have been developed. Such methods are PSQM [5], TOSQA [6] and the new PESQ [7]. A typical example of the correlation between the quality perceived subjectively and the predicted speech quality is shown in figure 5. The results are expressed in mean opinion scores where MOS 1 represents unacceptable quality and MOS 5 represents a very high quality.

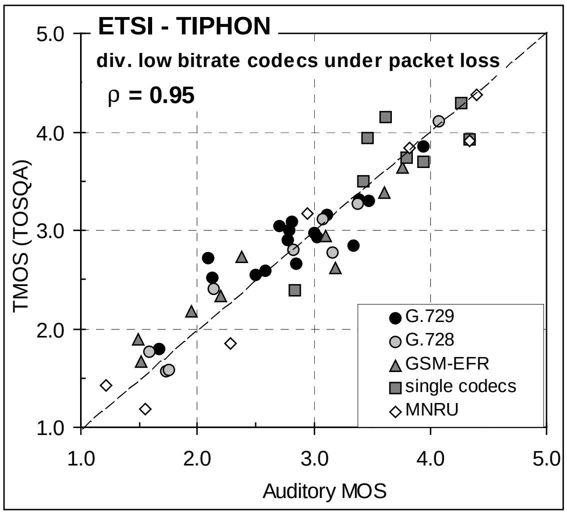

Fig. 5: Comparison of MOS values derived from auditory tests and MOS values predicted by TOSQA (TMOS) (see [6])

Those methods can be used for known codecs and for known impairments in order to evaluate speech sound quality. All those methods are perceptional based, but allow speech quality evaluations only in single talk conditions and unfortunately no method has been validated yet in order to evaluate the speech quality of systems including terminals.

Those methods can be used for known codecs and for known impairments in order to evaluate speech sound quality. All those methods are perceptional based, but allow speech quality evaluations only in single talk conditions and unfortunately no method has been validated yet in order to evaluate the speech quality of systems including terminals.

The transmission quality of background noise is - from the subjective point of view - one of the most important parameters. Since background noise is also transmitted during phases where no speech is present, background noise must be regarded as a signal. The transmitted background noise contains important information for the conversational partner about the environmental conditions. Echo cancellers, VAD's, DTX, and other algorithms are used in order to eliminate speech pauses/background noise in order to reduce bandwidth for transmission. Such operations may strongly degrade the quality of background noise transmission since pauses typically are filled by comfort noise which may interact with the transmitted background noise and lead to modulation of background noise.

D. Evaluation Methods and Examples

The most important evaluation methods for the dominating parameters influencing the speech quality in its various aspects are given using selected examples.

1.. *Background noise transmission quality*

Since the transmission of background noise subjectively is one of the most important parameters and in addition it is highly affected by the signal processing in IP connections the objective measurement of this parameter is of high importance as well. When analyzing the background noise transmission mostly the influence of the transmission system on the background noise is of importance, any structure inherent to the background noise itself should not influence the result of the analysis, such it is useful for analysis purposes to start with the evaluation of more or less constant background noises (e.g. office-type noises, street-type noises, interior vehicle noises for mobile terminals). The analysis methods used for that purpose is the "relative approach" [8]. The algorithm does not use any reference signal, but is working directly on the transmitted background noise signal. It takes into account the sensitivity of the human ear on a signal fluctuation in the time domain as well as on dominant spectral structures. It is recognized that slow variations of a signal in time and/or frequency are typically not disturbing which is taken into account by the algorithm. The algorithm is based on forward estimation using the signal history. The new signal examples are predicted and compared to the actual acquired signal. The basis for the procedure is the hearing adequate spectral representation of the time and frequency domain. The basis therefore is a hearing model according to [9]. The nonlinear relationship between sound pressure level and loudness perceived subjectively is taken into account by time/frequency warping in a Bark filter bank and proper integration of the individual outputs. The filter bank is realized in the time domain. The output signals of the filter bank are rectified and integrated, thus the envelope is generated. The three-dimensional output of the hearing model is the basis for the relative approach. In each critical band long term level (integration time: 2-4s) is compared to the short term level (2ms).

An overall value can be derived for example by applying the following equation (see [8]):

$$Q = f(N,S) + f\left(\sum_{i=1}^{24}\left[\left|F_G(i-1) - F_G(i)\right| \cdot w_1(iF_G(i)) + \sum_{n=1}^{T}\left|F_G(i,n) - F_G(i,n+1)\right| \cdot w_2(i,F_G(i))\right]\right)$$

where F_G (i) is a mean value of the critical band level over a period T of 2 to 4 seconds, F_G (0) = F_G (1), F_G (i, n) is a mean value of the critical band level over a much shorter period (approx. 2 msec), n is the current (time-dependent) value. The weighting factors $w_1(i,F_G(i))$, $w_2(i,F_G(i))$ depend on the critical band level F_G (i). In addition the overall value is influenced by the function f (N,S) which describes an audititory factor, dependent on loudness N and sharpness S.

When just displaying the result in the time/frequency domain a typical result looks like the picture below. The difference is color-coded and represented as a type of spectrography: high color represent big differences between estimation and actual signal, dark colors indicate low differences.

Fig. 6 shows the analysis result of a background noise reduction algorithm used in a hands-free terminal. The upper part of the picture shows the time response (light color) when switching on the hands-free terminal in the presence of background noise (the dark color indicates the time signal of the background noise signal).

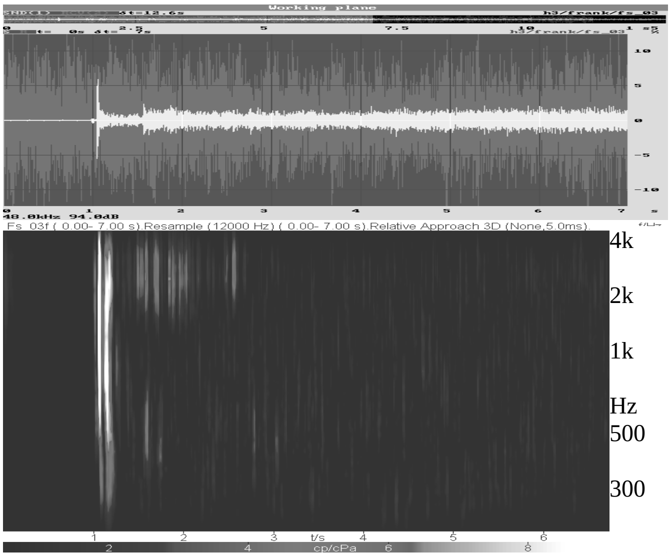

Fig. 6: Relative Approach Analysis,

upper: corresponding time signal
(light: transmitted background noise signal,
dark: background noise signal)
lower: Relative Approach analysis
(light: annoying signal components)

From the time signal it is not very obvious whether the algorithms works properly or not. When however analyzing the result of the relative approach it can be seen that the high peak in the beginning of the adaptation process is auditory annoying, the energy is spread over the whole frequency range for a small period of time. Within the first two seconds additional structures can be found between 1.6 and 3.4 kHz which are obvious in the relative approach and are obvious as well during the auditory judgement of the background noise transmission. Similar results could be achieved for other sorts of background noises and algorithms. Research work is ongoing in order to quantify the result of the relative approach analysis.

2.. Duplex performance

During double talk speech quality may be affected mainly by echo and/or switching. Special test signals an analysis techniques were developed in order to simulate the double talk behavior in the most realistic way but being able to analyze reliably and repeatable this situation. One test signal configuration used is shown in Fig. 7.

The signals are based on a CS-signal [10] simulating voiced/unvoiced sounds of speech, typical power density, modulation and distribution of speech using deterministic signals: voiced sound in the beginning followed by a pn-sequence and a pause. The double talk signal is constructed similar but uncorrelated to the test signal. During the pauses of the double talk signal -these are the time intervals where for real conversations echo and switching is audible- evaluations on echo attenuation and switching can be made.

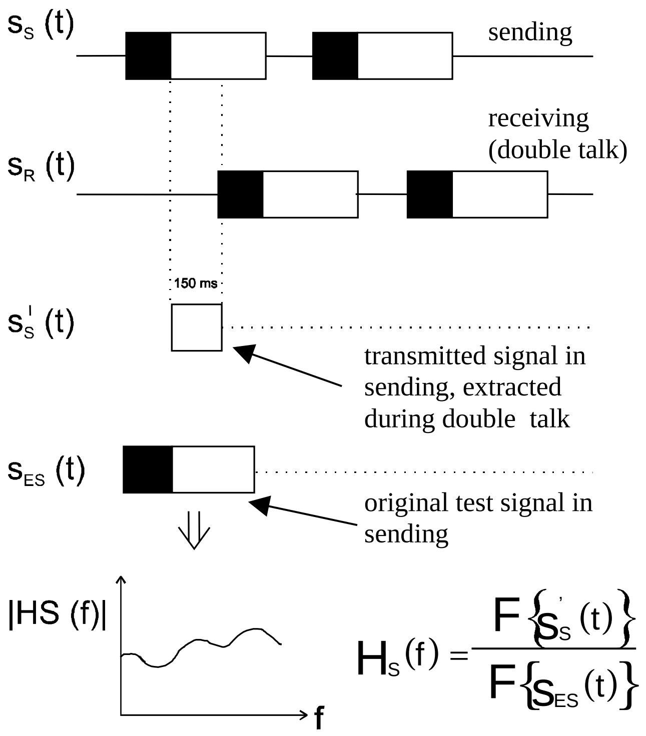

Fig. 7: Test signal for double talk evaluations

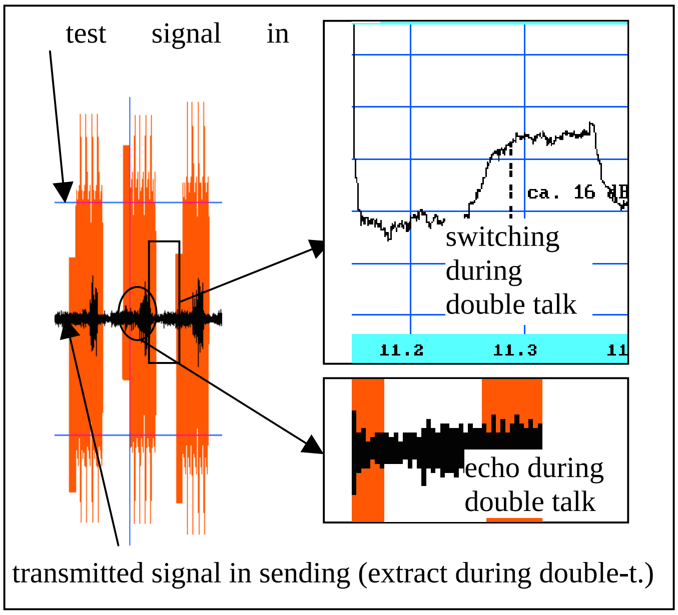

Fig.8: Example of extracting signal components for switching and echo evaluation during double talk in sending direction

Fig. 8 gives one example how to extract the information about switching and echo during double talk based on the test introduced with Fig. 7. Switching is evaluated during periods where only the signal in sending direction is present. The echo is determined during periods, where only the double talk signal is present and no signal in sending direction should be expected. In general all these investigations have to be conducted under various network conditions in order to get a range of performance requirements.

E. Summary

The present article discusses impairments which may occur in modern VoIP systems during speech communication. Starting from functional units typically found in connections a short overview about signal processing is given. Instrumental procedures which may be used for the investigation of the dominating speech quality parameters in VoIP systems are introduced. Methods and performance requirements are based on auditory tests and cover the conversational situation including the transmission of background noise. Especially for the double talk situation and the background noise transmission the evaluation procedures and the underlying principles namely the auditory test results are given.

F. Acknowledgements

Part of the work was conducted in project with T-Nova Deutsche Telekom Berkom.

G. References

[1] ITU-T Recommendation G.131

[2] F. Kettler; Gierlich, H.W.; Diedrich, E. Echo and Speech Level Variations During Double Talk Influencing Handsfree Telephones Transmission Quality, IWAENC 99, 27- 30.9.1999, Pocono Manor, USA

[3] ITU-T Recommendation P.340

[4] Sotscheck, J.: Über die Wahrnehmbarkeit von Clipping-Erscheinungen am Wortanfang, DAGA 90, pp. 1119-1122

[5] ITU-T Recommendation P.861

[6] Berger, J.: Instrumentelle Verfahren zur Qualitätsschätzung, Ph.D. Thesis, 1998, Shaker Verlag, ISBN 3-8265-4091-3

[7] Draft ITU-T Recommendation P.862

[8] Genuit, K. Objective Evaluation of Acoustic Quality Based on a Relative Approach, Internoise '96, Liverpool, UK

[9] Sottek, R.: Modelle zur Signalverarbeitung im menschlichen Gehör, PHD thesis RWTH Aachen, 1993

[10] Gierlich, H. W.: A Measurement Technique to Determine the Characteristics of Hands-Free Telephones, Signal Processing, Vol. 27, Issue 3, 1992