Speech Quality – A Multidimensional Problem: An Approach to Combine Different Quality Parameters

Hans Wilhelm Gierlich, Frank Kettler

HEAD acoustics GmbH, 52134 Herzogenrath, Germany, Email: H.W.Gierlich@head-acoustics.de

Summary

Modern telecommunication equipment is using a variety of non-linear and time variant signal processing techniques, found in both terminals and networks. The conversational quality of such equipment cannot be determined by a single value. A variety of non independent parameters contribute to the overall quality. Different approaches how to combine the different contributing parameters are discussed: objective speech quality measures like PESQ [1], TOSQA2001 [2], the E-model [3] and a combined approach using different speech quality parameters. Based on existing techniques the combined parameter approach seems to be the most promising one.

Signal Processing in Modern Networks

Traditional telephones and traditional transmission systems like switched or TDM networks are more and more replaced by IP based transmission, also modern mobile networks work on packed basis. The variety of terminals is increasing dramatically: very small mobile terminals, PDA type mobile terminals, mobile terminals including hands-free functionalities, various types of car hands-free terminals and a variety of computer based office phones are used instead of the traditional handset phones at many locations. In most of the modern terminals advanced signal processing is needed in order to guaranty a sufficient speech quality under the different conditions the terminal is used. Especially mobile terminals are used in much noisier environments than traditional office type terminals. The importance of the background noise and its transmission in mobile phones is much more significant than it was in the past. Figure 1 shows an example of the different signal processing blocks to be found in modern terminals.

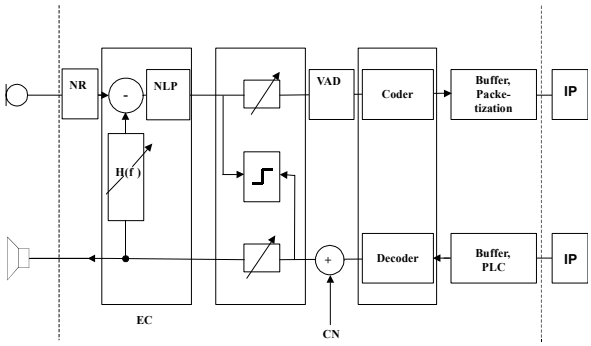

Fig. 1: typical signal processing in modern terminals

It can be seen that mostly all of the signal processing blocks are neither linear nor time invariant. In packed based networks a jitter buffer in combination with packet loss concealment takes care of the non-predictable behavior of the network concerning the delay of the individual packets. Therefore typical jitter buffers are adaptive (time variant delay). Packet loss concealment algorithms may reconstruct lost speech signal information (lost packets) using various strategies. The speech coders used are non-linear and time variant. Echo cancellation is needed in mostly all modern terminals. This is due to either the size of the terminals (strong coupling between microphone and receiver) and/or too long and unpredictable signal delays in the networks which leads to audible echoes in case that the echo attenuation is not sufficient. The performance of echo cancellers strongly depends on the echo path, the speech signals to be cancelled and the double talk signal as well as the background noise at the near end. All types of voice controlled attenuation or post-processing which may help to further reduce echo components are used.

Speech Quality from the Perceptual View

The perception of speech quality in modern telecommunication is complex and influenced by a variety of parameters. In general the communicational quality is representing the overall speech quality. Prom the perceptual point of view the communicational quality is influenced by:

- Listening Speech Quality
- Talking Speech Quality
- Speech Quality during Double Talk
- Background Noise Transmission Quality

For the assessment of the speech quality a variety of subjective testing methods are described in e.g. ITU-T ([4], [5], [6], [7]): Conversational Tests, Listening Tests, Talking and Listening Tests, Double Talk Tests, Speech Quality Tests in Background Noise situations.

Although the parameters listed above look independent they may influence each other. E.g. the listening speech quality may completely dominate the perception of the overall speech quality in case of high (bursty) packet loss introducing gaps and audible artifacts in the listening situation (see [8]). In conversational situations with high delay the double talk performance is of higher importance since double talk will occur more incidentally compared to situations with low delay. In situations with high background noise the quality of background noise transmission and speech quality in the presence of background noise are of much higher importance compared to quiet situations where these parameters may be neglected.

The Combination of Individual Objective Speech Quality Parameters

The measurement and prediction of the overall speech quality requires –independent of the methodology chosen- the combination of individual objective speech quality numbers. These are derived from testing methodologies validated by the different subjective testing procedures available.

Objective Speech Quality Measures

Although often used exclusively, speech quality measures like PESQ [1] or TOSQA2001 [2] are not able to predict the overall speech quality. The models compare the transmitted (degraded) speech signal to a reference speech signal. The signals

are time aligned and each is processed by a hearing model. The output signals of the hearing models are compared, weighted and mapped to a single value representing the MOS value derived in a listening speech quality tests. The methods can be used to predict the speech quality in the listening situation under some restrictions:

- The methods model the speech quality perception measured in listening tests using ACR rating
- The effect of different listening levels is not modeled
- The types of impairments (degradation e.g. due to speech coding) introduced by the technical transmission system must have been used in the validation process of the models
- The methods do not provide a relationship between the listening speech quality and the overall speech quality in the different communicational situations

The E-Model Approach

The E-model [3] is a network planning tool which tries to combine the different impairments contributing to the overall speech quality. The main assumption made by the E-model is that impairments used in the E-model are independent and additive on a psychological scale. The basic E-model approach is represented by the formula:

$R = R_0 - Is - Id - Ie + A;$

R_0 – basic S/N, Is – simultaneous impairment, Id – delayed impairment, Ie – equipment impairment, A – advantage factor.

For classic telephone networks and terminals the E-model approach was shown to be valid and very useful. Loudness Ratings, delays, echo attenuation with a reasonable degree of confidence were assumed to be time invariant. The influence of codec distortions could be modeled by the equipment impairment factor – to some extent even under the conditions of packet loss (see [3]). The current model is not suitable for the modeling: time variant behavior of systems (e.g. echo cancellers) or different system behavior under different conversational conditions (e.g. double talk). The system performance under background noise conditions cannot be modeled at all. It can be assumed that the general approach taken by the E-model namely the independency and the additivity of impairment factors cannot be kept in the future and some general modifications of the model for non LTI-systems is required. Furthermore it has to be considered whether one rating factor is really suitable to represent the different conditions modern (terminal) equipment is used. A general case decision, based on different environmental conditions and depending on different use cases may be a suitable way forward. Within the individual impairment factors a different weighting scheme should be investigated. This weighting scheme must take account the dependency on the different impairment factors in the different conversational situations.

The Combined Parameter Approach

The combined parameter approach in general is similar to the E-model approach however the aim is not to determine a single value representing the speech quality. The combined parameter approach uses a graphical representation of the contributing speech quality parameters. For a terminal the following parameters may be taken into account:

- SLR, RLR and MOS-LQO in sending and receiving
- Delay and TCLw
- Double talk performance based on the classification scheme developed for hands-free terminals (see [9] and [10])
- Performance with background noise e.g. based on the Relative Approach but new methods for determining the speech quality in the presence of high background noise are required

For networks a similar selection of parameters may be used. One example for a possible representation of parameters in one diagram is shown in Fig. 2. Each axes represents one parameter. All axes are scaled such that quality improvements lead to higher distances from the origin. Symmetric parameters (e. g. Loudness Ratings) can be represented by reversing scales. Such the minimum requirements can be adapted easily. Although not a single value is created (this might be possible for specific conditions where the relationship between the parameters and their contribution the overall quality is known better) this representation gives visually an excellent overview over the speech quality and the contribution of the individual parameters. Impairments as perceived subjectively can be identified immediately which allows more easy to refer those to problems in the individual technical implementation.

Further work is needed in order to better describe the impairments in background noise situation (background noise reduction and their effect on speech quality and noise transmission performance) and to better understand the relation between the parameters and their individual contribution to overall quality.

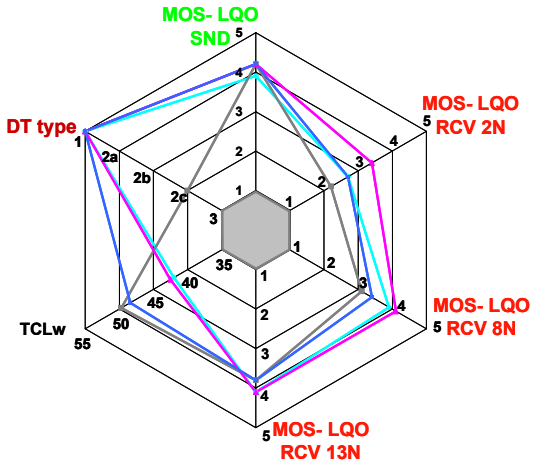

Fig. 2: Example for a combined parameter representation, 4 phones (A: light blue, B: grey, C: magenta, D: blue) see [11]

References

[1] ITU-T Recommendation P.862
[2] Berger, J. : Results of objective speech quality assessment including receiving terminals using the advanced TOSQA2001, ITU-T Contribution, Dec. 2000, COM 12-20-E
[3] ITU-T Recommendation G.107
[4] ITU-T Recommendation P.800
[5] ITU-T Recommendation P.831
[6] ITU-T Recommendation P.832
[7] ITU-T Recommendation P.835
[8] Raake, A.: E-Model: Additivity of burst packet loss impairment with other impairment types, ITU-T SG12 2004, D. 221
[9] ITU-T Recommendation P.340
[10] Gierlich e.al.: Proposal for the Definition of Different Types of Hands-free Telephones Based on Double Talk Performance, ITU-T SG12, Sept. 99, COM 12-103
[11] Kettler e.al.: Speech Quality for VoIP Terminals, DAGA 04