Speech-Quality Evaluation in Telephone Networks

Thorsten Ludwig, Kirstin Scholz, Ulrich Heute
Institute for Circuits and Systems Theory, University of Kiel; Email: {tlu,ks,uh}@tf.uni-kiel.de

Introduction

The technical progress in the past 15 years, resulting in new services and devices for telephone networks, has formed a telecommunication infrastructure that is characterized by an almost infinite number of possible network configurations. Simulations of telephone networks in laboratories can model real effects in networks only to a limited extent. Subjective tests for the evaluation of speech quality are time-consuming and expensive. For the assessment of speech quality in real-life telephone networks other methods are necessary. One approach to investigate real-life telephone links are measurements in the telephone network by means of in-service, non-intrusive measurement devices (INMD). These devices can serve as network monitors that evaluate the quality-of-service parameters of the network under test. A high correlation with subjectively determined quality values is a prerequisite for the use of these devices. This paper gives an overview of signal parameters that are useful for speech quality assessments by means of INMDs. Some algorithms to measure these parameters are presented. For the evaluation, measurement results of an INMD can be used as input parameters to network planning models to get estimations about the speech quality.

In-Service, Non-Intrusive Measurement Device

The monitoring of quality-of-service parameters of a telphone network can be done by in-service, non-intrusive measurement devices (INMDs) [1,2]. Therefore, INMDs are placed at switches (on PCM-coded lines) in the network. They observe a multitude of telephone calls during network operation (in-service, non-intrusive). A central evaluation gives evidence about network anomalies or the transmitted speech quality up to the measuring point. The quality evaluation is often based on network planinng models (e.g. the E-Model [3]), where INMD-measurements are mapped to input parameters of these models.

Sources for Speech-Quality Degradations in Telephone Networks

Many necessary network components are at the same time sources for degradations in the telephone network. Digital circuit multiplication equipment (DCME) devices, for example, make use of codecs with lower transmission rates compared to the standard PCM in order to increase the channel-capacity of telephone channels. Additionally, voice activity detectors (VAD) blank speech pauses to save bandwidth. Both, the cascading of codecs in DCMEs and the blanking of speech pauses with the inherent clipping of speech fragments, result in speech quality degradations. Echocancellers influence the duplex-communication functionality of a telephone link. The standardized codecs for speech transmission offer different speech transmission qualities. Noise reduction systems interwork with other components resulting in further degradations. The radio channel in mobile communications causes bit-errors and frame losses. Voive over Internet Protocol (VoIP) calls linked to a subscriber in the PSTN via a gateway suffer from packet losses caused by long signal delays and delay jitter.

Traditional parameters as speech level, noise level, echo loss, and echo delay are no longer sufficient to describe the speech quality of telephone links. Therefore, the measuring of additional parameters is necessary. Examples are the detection of comfort noise, of digital transmission systems (codecs) [4,5,6], of doubletalk, and of frame/packet losses. The next section presents a short overview about the detection of frame loss in mobile communications and packet loss in VoIP-calls in case of basic substitution techniques.

Frame-Loss (Mobile Communication) / Packet-Loss (VoIP)

Frame Loss occurs due to channel fading, shadowing and multipath fading on the radio channel. The basic substitution technique for lost frames repeats the decoding of the last received parameters. As a result, adjacent frames are highly correlated. After several lost frames the call is muted leading to regions of very low energy in the signal. For the detection of lost frames, the correlation $\rho_{i,\tau}$ of adjacent frames y(k,i) and y(k,i+τ) is calculated:

$$\rho_{i,\tau} = \frac{\sum_{k=0}^{M-1} y(k,i)\cdot y(k,i+\tau)}{\sqrt{\sum_{k=0}^{M-1} y^2(k,i)}\cdot\sqrt{\sum_{k=0}^{M-1} y^2(k,i+\tau)}}, \quad (\Delta M = 0 \rightarrow \tau = 1) \tag{1}$$

$$y(k,i) = y(k + i\cdot(M-\Delta M)), \quad M = 160$$

M describes the block length, ΔM the overlapping of signal-blocks. Two problems occur. First, a high correlation may indicate not only a lost frame, but also a speech-signal fundamental frequency f_0 which is a multiple of the frame rate (50 Hz for GSM). Therefore, it has to be checked that f_0 is not a multiple of 50Hz in the range of typical fundamental frequencies (100Hz-400Hz) [6]:

$$f_0 \neq \eta\cdot 50\text{Hz}, \quad \eta = 2,3,\ldots,8 \tag{2}$$

or, as the practical estimation of f_0 will never yield exact values, according to equation (2):

$$f_0 \notin \left[\eta\cdot 50\text{Hz} - \varepsilon(\eta), \eta\cdot 50\text{Hz} + \varepsilon(\eta)\right], \quad \eta = 2,3,\ldots,8 \tag{3}$$

with ε(η) defining a small range around the multiples of 50Hz. A second problem is the lack of knowledge about the beginning and ending points in time of the GSM-frames. To be exact, the correlation analysis must be performed for every signal value y(k) as a possible startpoint of a GSM-frame. Further investigations show that it is sufficient to use every fourth point in time:

$$\rho_{i,\tau} = \frac{\sum_{k=0}^{M-1} y(k,i)\cdot y(k,i+\tau)}{\sqrt{\sum_{k=0}^{M-1} y^2(k,i)}\cdot\sqrt{\sum_{k=0}^{M-1} y^2(k,i+\tau)}}, \quad (\Delta M = 156 \rightarrow \tau = 40) \tag{4}$$

In VoIP-calls the basic substitution algorithms depend on the same principles as in mobile communications. Figure 1 shows as example an undisturbed signal clip compared to the same signal clip after substitution of lost packets. Clearly visible are phase jumps at signal block boundaries and the muting after several lost packets. Muted signal parts or inserted zeros substituting losses can be detected by evaluating the energy gradient of short signal blocks (5ms). In the following this is done in the frequency domain:

$$G(i,i+1) = \sum_{\mu=\mu_l}^{\mu_u} \left|Y(\mu,i+1)\right|^2 - \left|Y(\mu,i)\right|^2 \tag{5}$$

High ($\mu > \mu_u$) and low frequencies ($\mu < \mu_l$) are disregarded. A realization in the time domain is also possible. Y(μ,i) is the DFT of

the i-th signal block y(k,i) with M=40 and ΔM=0. G(i,i+1) is the energy gradient for two adjacent signal blocks. After normalization we get:

$$G^n(i,i+1) = \min\left(\frac{G(i,i+1)}{\sum_{\mu=\mu_l}^{\mu_u} |Y(\mu,i)|^2}, 1\right), \quad -1 \le G^n \le 1 \qquad (6)$$

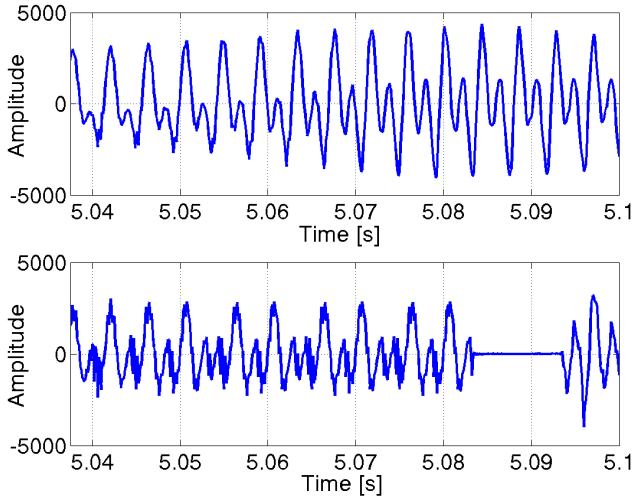

Figure 1: G.711 packet loss. Upper: original signal, lower: signal after substitution of lost packets (packet length 10ms).

Figure 2 shows an example. At the beginning of each muted signal part the energy gradient is $G^n(i,i+1)=-1$. When the signal is present again the gradient jumps to $G^n(i,i+1)=1$.

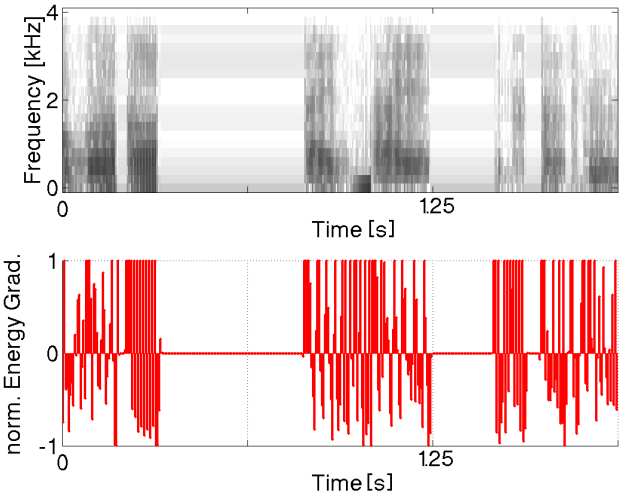

Figure 2: Upper: spectrogram of signal clip, lower: normalized energy gradient $G^n(i,i+1)$.

Quality Evaluation

For the quality evaluation, signal parameters measured by an INMD can serve as input parameters to network planning models. In the following Example, 44 GSM-FR signals are investigated. The signals are corrupted by different background noises (street noise, office noise, in-car noise and low level noise), bit-errors and frame losses. Two signals showed divergences from the standard speech level. The measured quantities for the quality evaluation are the speech level, the noise level, the detection of the GSM-FR-codec and the detection of frame losses. The measured parameters are transformed to input parameters of the E-Model according to [2]. For frame loss a transformation rule described in [7] was used, as there exists no standardized method in [2]. Not measureable input parameters of the E-Model are set to standard values for telephone networks. The E-Model results are transformed to mean opinion scores (MOS) as used in subjective listening tests. Figure 3 shows the result. There is a high correlation of ρ=0.965 between the subjective determined MOS-values and the instrumentally derived MOS-values. The standard deviation is σ=0.283.

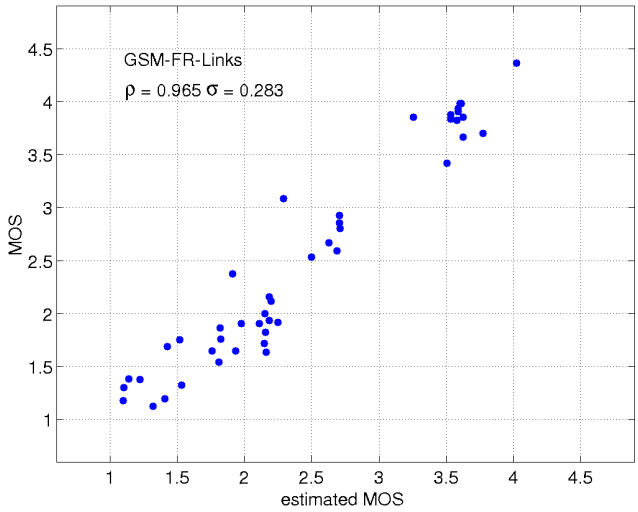

Figure 3: Speech quality evaluation; subjectiv MOS over estimated MOS (E-Model).

Conclusions

The paper describes the use of in-service, non-intrusive measurement devices for the automatical speech-quality monitoring and evaluation in telephone networks. The evolution of the telephone network has led to many new sources for disturbances. Therefore, besides standard parameters, new quantities have to be measured. As examples, the measuring of frame and packet loss in case of basic substitution principles for lost signal parts are presented. The last part describes the E-Model as one possibility for the evaluation of quantities measured by an INMD. The result shows a high correlation between subjective MOS-values and instrumentally derived MOS-values based on INMD measurements.

References

[1] ITU-T Recommendation P.561, In-Service, Non-Intrusive Measurement Device – Voice Service Measurements, 1996.

[2] ITU-T Recommendation P.562, Analysis and Interpretation of INMD Voice-Service Measurements, 2000.

[3] ITU-T Recommendation G.107, The E-Model, a Computational Model for Use in Transmission Planning, 1998.

[4] Th. Ludwig, U. Heute, Detection of Digital Transmission Systems for Voice Quality Measurements, EUROSPEECH, 2001.

[5] Th. Ludwig, Comfort Noise Detection and GSM-FR-Codec Detection for Speech-Quality Evaluations in Telephone Networks, ICSLP, 2002.

[6] C. Veaux, P. Scalabert, A. Gilloire, Analysis and On-Line Detection of Audible Distortions in GSM Telephony, EUROSPEECH, 1999.

[7] S. Möller, A. Raake, Telephone speech quality prediction: Towards network planning and monitoring models for modern network scenarios, Speech Communication 38, 2002.