6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

QUALIPHONE-A: A PERCEPTUAL SPEECH QUALITY EVALUATION SYSTEM FOR ANALOG MOBILE NETWORKS

M. Szarvas, T. Fegyó, P. Tatai and G. Gordos

TSP Laboratory, Department of Telecommunication and Telematics,
Technical University of Budapest
Pázmány Péter sétány 1/D, 1117 Budapest, Hungary
E-mail: {szarvas, fegyo, tatai}@bme-tel.ttt.bme.hu

10.21437/Eurospeech.1999-563

ABSTRACT

This paper describes an objective speech quality assessment method developed for the Hungarian NMT-450 mobile telephone system. The method is based on a psychoacoustic front end followed by a cognitive modeling component. Special problems of the NMT system, such as hand-overs, the effects of automatic gain control (AGC) and intrusion of signaling noise are addressed in the cognitive module. Correlation of the subjective and objective quality measures is maximized by finding a transformation that linearizes their relationship. A correlation of 0.94 is achieved on an independent test set between the subjective speech quality and the proposed objective quality measure.

Keywords: *speech quality estimation, analog mobile telephone, psychoacoustic modeling*

1. INTRODUCTION

Monitoring the quality of speech communication channels is essential for ensuring the required quality of service. It is particularly critical in analog mobile environments where a large number of factors can result in degradations. Unfortunately, neither conventional analog measurements nor perceptual quality estimation methods developed for low bit-rate voice codecs could be used directly in the analog mobile environment because

- the speech transmission channel is difficult to characterize or model due to its nonlinear and time varying nature,
- hand-overs result in short periods of silence or signaling tone transmissions during conversation which must be considered as normal phenomena,
- unpredictable white background noise, impulsive noise hits, interference, and fading effects are present in addition to linear and nonlinear distortions,
- most of the time the users can tolerate certain types of large degradations (especially hand-overs (HOV) and non-linear distortions) as long as the intelligibility is acceptable

Considering the problems above, an objective perceptual speech quality estimation method had to be developed which predicts fairly well the subjective quality of the analog radio telephone channels.

The main stages of processing in our system are the synchronization of the reference and the test signal, transformation to a psychoacoustic representation, cogntitve modeling, distance calculation and non-linear regression to maximize the correlation with subjective quality measurement results. The basic structure of the system is similar to those described in [1, 2, 3], except for the details of the applied psychoacoustic and cognitive processing.

2. SYNCHRONIZATION PROCEDURE

The first step of processing in our system is to time-align the reference and the test waveform data. (The waveform data to be evaluated is referred to as *test waveform.*) Ideally this task can be solved by the use of a quick speech activity detector followed by a correlation based aligner. However, during real world measurements this method is not always succesful in its most simple form. It may, and does indeed, occur that the portion of the signal we are looking for is missing from the recording due to a handover. In order to assure that the alignment signal is not missing, Qualiphone-A first selects the loudest part of the test waveform as the alignment signal. Because hand-overs have very little energy, this procedure avoids using them for synchronization. To ensure best results in later stages, the accuracy of the synchronization is 1 sample (10e-4 s).

3. PSYCHOACOUSTIC MODELING

The psychoacoustic layer utilizes an FFT based mel-filter bank, followed by components for modeling forward and backward masking, internal noise and compensation for linear distortions. The cognitive modeling layer performs special treatment of the hand-over regions, silence regions, signaling noise, and detection of premature hang-ups.

The steps of transforming to the inner representation are depicted graphically in Figure 1. and described in detail in the following:

1 One sample precision synchronization of the reference and the test signal.

2 Preemphasis. The signal is passed through two cascaded first order high pass FIR filters to compensate for the low pass filtering characteristics of the telephone channel. The pre-emphasis coefficient of both filters is 0.8.

3 Splitting the signal into 30 ms long frames with 20 ms overlap between the frames.

4 Energy normalization. Each sample of the frame is scaled by a frame dependent factor, so that the energy of each frame is the same after scaling. This step is necessary to compensate for the effects of fading and automatic gain control.

5 Decomposition into a 33 channel mel scale filter bank representation. The optimal dimensionality of the filter bank has been determined by experimentation. The filter bank is implemented by the FFT method. Each dimension of the output vector from this step represents the energy in the corresponding channel of the filter bank.

6 Detection of synthetic signals. Sometimes short periods of synthetic signals can be heard during the connection. These periods, similarly to HOVs and silence periods, should be treated separately from normal periods. Currently only the signal indicating premature hang-up is detected. The detection is based on measuring the percentage of signal energy in certain filter bands. There is a lower limit on the length of hang-up indicator. If a hang-up indicator is found, both the frames of the indicator and the frames following it are ignored during further processing.

7 Detection of hand-overs. Hand-overs (HOV) are regions of the signal where a change of base station happened. Unlike GSM, in NMT systems there is a small period of silence during hand-overs. Since hand-overs are natural during normal operation of NMT systems, the quality of the connection must not be considered low, just because the spectral difference is large during a handover. From a system design point of view, only the number of HOVs during a given unit of time is important, not the spectral difference during a HOV. Hand-overs are detected by using an energy threshold. If the energy in the test signal is lower than this threshold, and at the same time, it is lower than 30% of the energy in the reference signal, the frame is marked as a frame with HOV. After tagging individual frames, the adjoining frames tagged as hand-overs are joined in a hand-over region. Then, the length of each hand-over region is checked. If the length of the hand-over region is under a threshold, the HOV tag is removed from the frames, because HOVs can not be arbitrarily short. Finally, the non-HOV frames at the boundaries of HOV regions are also labeled as part of the HOV, because in the majority of cases, signaling noise can be heard in these frames, that results in a large spectral difference, but does not decrease the subjectively perceived quality, since human evaluators consider these frames as part of the HOV.

8 Silence detection. Regions of the reference signal with energy under a given threshold are identified. This is useful, because noise is not as disturbing during periods of silence as it is during periods of speech. Silence detection is based exclusively on energy thresholding. A silence region can be arbitrarily short, there is no length limit used here, as in the case of HOVs.

9 Assigning weights to different types of signal regions. The HOV, silence and synthetic signal regions are assigned a weight of 0.0, the rest of the frames are assigned a weight of 1.0. Although the experimental results in [1] indicated that the assignment of a non-zero weight to silence frames was optimal, in our experiments the 0.0 value seemed the best.

10 Smoothing of filter outputs. The output of each filter is smoothed with a 3 point FIR filter uniformly weighting every point. This step, followed by a non-linear transformation in step 12, is expected to model masking that exists in the human auditory system.

11 Injection of internal noise. In certain time-frequency regions of the signal the energy is close to zero. Therefore very small absolute differences of energies would cause large differences of their logarithms in a succeeding step, contrary to their small perceptual significance. (Small differences can not be heard even in silent regions, because of the inner noises in the human auditory system.) To avoid this unjustified difference, the internal noise of the auditory system is modeled by adding a channel and signal dependent amount of "noise" to each frame of every channel. The energy of the noise is the maximum of two quantities: it's either 0.3 or 10% of the average signal energy in the given channel.

12 Transformation to the logarithmic domain. During this step, the mel filter energies are replaced by their logarithms.

13 Compensation for linear distortion. This step is implemented by high pass filtering of filter outputs.

The (weighted) average during the preceding 500 ms of the logarithmic filter energy is subtracted from each frame of logarithmic filter output.

4. DISTANCE CALCULATION

The next step is to determine the average distance of the inner representation of the reference and that of the test signal. This distance is defined as the weighted average of the Euclidean distance of the modified filter-bank outputs of each reference- and test-frame. The weights used during averaging are those described in step 9. above. The distance determined this way is denoted by D_{inn} in the following.

The basic indicator of the speech quality is this D_{inn} distance. The relationship between D_{inn} and the subjective MOS values is monotonous and the correlation is relatively high (on the order of 0.75). The scale of the two quantities (D_{inn} and MOS) is, however, different. The other problem is that the relationship between the MOS values and D_{inn} is non-linear.

4.1 Linearization

In order to increase the correlation of the objective quality measure and the subjective MOS values, the next step is to find a transformation that makes their relationship linear (as much linear as possible). After investigating different types of transformations (different degrees of polynomials and the logarithm function) the logarithm of the D_{inn} distance was found to be most correlated with the MOS values. A shortcoming of the logarithmic transformation is that the resulting D_{lin} distance is not sufficiently correlated at the top of the quality scale (MOS=4.0-5.0) with the MOS values. Nevertheless, this transformation resulted in the best overall correlation, therefore this one is used in our system currently. The linearized distance described above will be referred to as D_{lin} in the following. To summarize, the relationship between the basic inner distance and the linearized distance is $D_{lin} = \log(D_{inn})$.

4.2 Scale transformation

The basic measure of goodness of a given objective quality assessment method is the correlation between the objective and the subjective scores. This correlation is maximized during the steps of transforming to the inner representation and by the linearization transformation. However, to be of practical use, the D_{lin} measure must be converted to the scale of the MOS values, that is, to the range of 0.0-5.0. (This range is used by the Hungarian NMT 450 service provider, as opposed to the 1.0-5.0 range in the ITU recommendation [4, 5]. The meaning of the 0.0 value is "unintelligible".) The relationship between the linearized distance and the final QualiPhone Objective quality Score is defined by the following linear relationship:

$$QPOS = a\, D_{lin} + b$$

The values of the constants, *a* and *b*, are determined during an optimization procedure. The criterion for the optimality is the minimization of the squared average distance of the *QPOS* values and the MOS values on the calibration set.

The scale transformation does not influence the correlation between the subjective and objective quality values. Its role is to help interpret the objective quality measures.

5. EXPERIMENTAL EVALUATION

A set of controlled experiments have been carried out in order to assess the quantitative relevance of each of the above signal processing stages. 125 sentences were selected and evaluated by a panel of 13 human listeners. The length of each sentence was approximately 4 seconds. The sentences were selected to cover the whole quality range approximately uniformly. The listeners were given detailed instructions based on the previous quality assessment standard applied at the service provider. An expert checked all of the scores and removed the obviously erroneous ones (outliers). After removing the outliers, the scores were averaged resulting in the Mean Opinion Score (MOS) for each sentence. 10 of the sentences were selected to form the calibration data set for estimating the optimal scale transformation. The calibration set covers the entire quality range and the variance of the subjective scores for each of the calibration samples is very small (not more then two opinions differ from the majority). The remaining 115 sentences were used for measuring the correlation and the mean squared difference of the MOS and the *QPOS* values.

Tables 1 and 2 display the results of experiments assessing the importance of the use of internal noise, smoothing of filter outputs and compensation for linear distortions. Table 2 displays the results for the case with explicit treatment of HOV regions, while Table 1 stands for the case of no special treatment of HOV's. Both tables confirm that all 3 techniques contribute positively to the correlation of the MOS and the *QPOS* scores, although the effects of smoothing are ambiguous. Nevertheless, the best results were achieved with the use of all three techniques, therefore our system is applying all of them.

The relationship between the subjective and objective quality scores is illustrated graphically in Figure 2. A more detailed analysis of the different psychoacoustic stages is described in [6].

6. CONCLUSION

An objective speech quality estimation system has been developed for use on analog radio-telephone channels. The system is based on a mel-scale filterbank followed by components modeling psychoacoustical and cognitive phenomena. The use of the proposed psychoacoustic processing increased the correlation between the subjective and objective results to 0.9472 from the baseline 0.8790 of the filterbank alone. This accuracy makes the system an ideal alternative to human evaluation for regular system monitoring, and is indeed used regularly in the Hungarian NMT system.

7. ACKNOWLEDGEMENTS

The research described in this paper has been supported by Westel Ltd, the service provider on the Hungarian NMT 450 MHz mobile-phone network. We are especially greatful for the useful ideas contributed by Pál Erdõs and István Szabó.

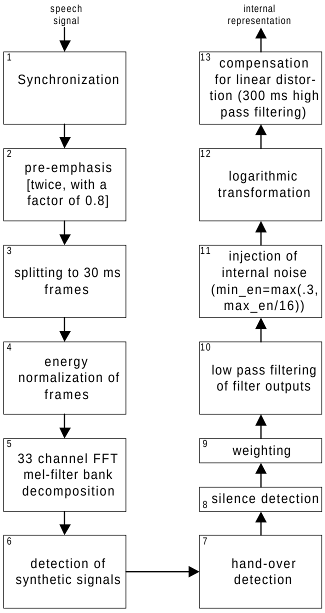
speech signal
1 Synchronization
2 pre-emphasis [twice, with a factor of 0.8]
3 splitting to 30 ms frames
4 energy normalization of frames
5 33 channel FFT mel-filter bank decomposition
6 detection of synthetic signals
7 hand-over detection
8 silence detection
9 weighting
10 low pass filtering of filter outputs
11 injection of internal noise (min_en=max(.3, max_en/16))
12 logarithmic transformation
13 compensation for linear distor-tion (300 ms high pass filtering)
internal representation

Figure 1: The steps of obtaining the internal representation from the time domain speech signal

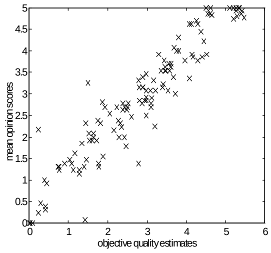

Figure 2: The relationship between the subjective and objective quality scores

Techniques used	correlation
none	0.8790
linear compensation	0.9316
internal noise added	0.8924
smoothing	0.8636
int. noise, smoothing	0.8771
internal noise and lin. comp.	0.9302
smoothing, lin. comp.	0.9353
int. noise, smoothing, lin. comp.	0.9332

Table 1: The effects of different psychoacoustical modeling techniques on the relationship between the subjective and objective quality measures when hand-overs are not treated specially.

Techniques used	Correlation
none	0.9317
linear compensation	0.9365
internal noise added	0.9281
smoothing	0.9268
int.noise, smoothing	0.9233
int.noise and lin.comp.	0.9447
smoothing, lin.comp.	0.9417
int. noise, smoothing, lin.comp.	0.9472

Table 2: The effects of different psychoacoustical modeling techniques on the relationship between the subjective and objective quality measures when hand-overs are detected and ignored.

8. REFERENCES

[1] J. G. Beerends, “Audio Quality Determination Based on Perceptual Measurement Techniques”, chapter 1 of *Applications of Digital Signal Processing to Audio and Acoustics,* Eds. M. Kahrs and K. Brandenburg

[2] “Method for Objective Measurement of Perceived Audio Quality”, *Draft New Recommendation ITU-R BS. [Doc. 10/20]*

[3] “Objective Quality Measurement of Telephone Band (300-3400 Hz) Speech Codecs”, *ITU-T Recommendation P. 861.*

[4] “Methods for Subjective Determination of Transmission Quality”, *ITU-T Recommendation P. 800. COM 12-65-E Jan. 96.*

[5] “Subjective Performance Assessment of Telephone-band and wideband digital codecs”, *ITU-T Recommendation P. 830. 02/96.*

[6] M. Szarvas, T. Fegyó, P. Tatai, G. Gordos, P. Erdõs, I. Szabó, “An Objective Speech Quality Estimation Method for Analog Mobile Telephone Channels”, in Proc. COST 254 Workshop on Intelligent Communication Technologies and Applications, with Emphasis on Mobile Communications, 1999.