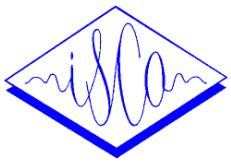

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

OBJECTIVE ANALYSIS OF THE GSM HALF RATE SPEECH CODEC CANDIDATES

F. Wuppermann* C. Antweiler** M. Kappelan**

*Philips Research Laboratories
P.O. Box 80.000, 5600 JA Eindhoven, The Netherlands
** Institute for Communication Systems and Data Processing
Aachen University of Technology
Templergraben 55, 52056 Aachen, Germany

ABSTRACT

During the standardization process of the GSM half rate codec, extensive subjective listening tests were performed in November 1992 in order to select the final half rate codec candidate for the GSM system. In addition an unofficial objective analysis was performed by the authors using an objective measure for speech quality.
This paper describes the fundamentals of this objective quality measure based on psychoacoustics, see also [1],[2]. A detailed comparison is given between the results of the subjective listening tests in terms of Mean Opinion Score (MOS) and the objective quality measure. A high correlation between both measures was observed for slightly distorted speech, a poorer correlation for larger distorted speech.
Keywords: *objective quality measure, GSM half rate codec selection test*

1. INTRODUCTION

Up to now, extensive subjective listening tests are performed to reliably judge the quality of processed speech signals (e.g. [3]). The results of these tests are expressed e.g. in terms of mean opinion scores (MOS), a one dimensional scale from 1 (bad quality) to 5 (good quality). An objective measure of speech quality, which is capable to predict MOS scores of processed speech signals, must fulfill two requirements:

1. The model must indicate whether distortions in the processed signal are audible.
2. If distortions are audible, the model must give an indication how the distortion contributes to the overall perceived quality of the processed signal.

Especially the second requirement is difficult to realize. To fulfill the first requirement we consider present day low-bit rate audio codecs (e.g. [4]). They use a psychoacoustic model to quantize audio signals in such a way that the quantization noise is masked by the signal. This model can also be used to provide an indication about the audibility of distortions. An advanced psychoacoustic model is described in [5]. In this model multiple critical band noises are masked if their excitation in each critical band [6] is just below a threshold. This threshold can be interpreted as the masked threshold.
Assuming that the distortion in processed speech signals can be considered as multiple critical band noises, we derived an objective quality measure (OQM) of speech from [5]. The main parts of this quality measure are described in Section 2.
Within the scope of the standardization process of the final GSM half rate codec extensive subjective listening tests were performed in November 1992. In addition an unofficial objective analysis was performed applying the derived objective quality measure (OQM). The test procedure used for the objective quality measurement is described in Section 3. A detailed comparison between the results of the subjective listening tests in terms of Mean Opinion Score (MOS) and the objective quality measure is provided in Section 4. Finally Section 5 concludes the paper.

2. OQM MODEL

In Fig. 1 a block diagram of the OQM model is given. The OQM model needs two signals: a speech signal with reference quality, called reference signal, and a distorted speech signal, called processed signal. Both signals have a sampling frequency of 8 kHz. It is required that there is no delay between the reference and the processed signal.

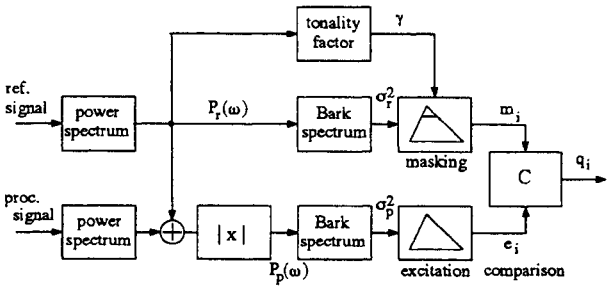

Figure 1: Block diagram of OQM model

First we consider the reference signal. Every 24 ms (192 samples) the reference signal is windowed by a centered Hanning window of 256 samples (32 ms). By an FFT of 256 points (frequency resolution 31.25 Hz), the signal is then transformed into the frequency domain where the power spectrum $P_r(\omega)$ is calculated.
Based on $P_r(\omega)$, a Spectral Flatness Measure (SFM) is used to calculate a tonality factor γ [7]. The tonality factor γ is in the range of 0 to 1; $\gamma = 0$ denotes a noiselike signal, $\gamma = 1$ denotes a tonal signal.

In order to derive a Bark spectrum $\sigma_r^2[j]$, the power spectrum $P_r(\omega)$ is divided into 17 critical bands and summed up within each critical band:

$$\sigma_r^2[j] = \sum_{\omega=\omega_{l_j}}^{\omega_{u_j}} P_r(\omega) \qquad j = 1, ..., 17, \tag{1}$$

where ω_{l_j} and ω_{u_j} denote the lower and upper bound of the critical band j, respectively. For a noiselike masker with power $\sigma_r^2[j]$ in critical band j, $1 \leq j \leq 17$ we define the masking pattern as function of i

$$m_n(\sigma_r^2[j], i, j) = \begin{cases} \frac{1}{10}\,\sigma_r^2[j]\beta^{|j-i|} & i < j \\ \frac{1}{10}\,\sigma_r^2[j] & i = j \\ \frac{1}{10}\,\sigma_r^2[j]\alpha^{|j-i|} & i > j \end{cases} \tag{2}$$

and for a tonal masker as

$$m_t(\sigma_r^2[j], i, j) = \begin{cases} \frac{1}{10}\,\sigma_r^2[j]\beta^{|j-i|} & i < j \\ \frac{1}{1000}\,\sigma_r^2[j] & i = j, j+1 \\ \frac{1}{10}\,\sigma_r^2[j]\alpha^{|j-i|} & i > j+1. \end{cases} \tag{3}$$

The parameters α and β are chosen such that the slope towards lower critical bands is 25 dB/Bark and towards higher critical bands is 10 dB/Bark [8]. Both masking patterns are illustrated in Fig. 2 on a Decibel-Bark scale for $\sigma_r^2[j] = 60$ dB.

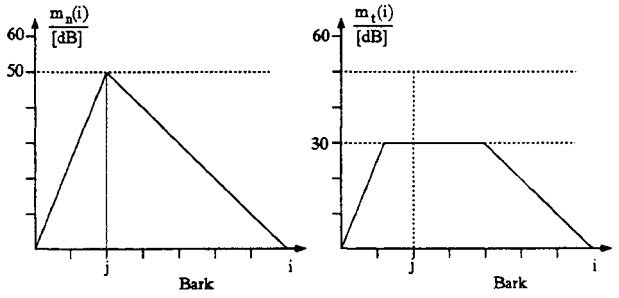

Figure 2: Masking patterns for a noiselike signal (left) and a tonal signal (right) for $\sigma_r^2[j] = 60$ dB

In order to determine the masked threshold m_i in critical band i caused by multiple maskers, we assume that masking is additive

$$m_i = \sum_{j=1}^{17} (\gamma m_t(\sigma_r^2[j], i, j) + (1-\gamma) m_n(\sigma_r^2[j], i, j) + t_i) \tag{4}$$

where t_i denotes the threshold in quiet in critical band i.

In the next step the absolute difference $P_p(\omega)$ between the power spectrum of the reference signal and the power spectrum of the processed signal is determined. Based on $P_p(\omega)$ we calculate the Bark spectrum $\sigma_p^2[j]$. Analogous to the determination of the masked threshold, the expression

$$e(\sigma_p^2[j], i, j) = \begin{cases} \sigma_p^2[j]\beta^{|j-i|} & i < j \\ \sigma_p^2[j] & i = j \\ \sigma_p^2[j]\alpha^{|j-i|} & i > j \end{cases} \tag{5}$$

denotes the excitation in critical band i caused by a signal with power $\sigma_p^2[j]$ in critical band j. This excitation pattern is illustrated in Fig. 3 on a Decibel-Bark scale.

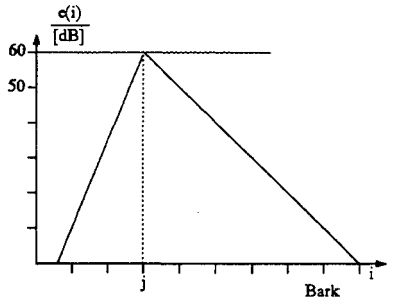

Figure 3: Excitation pattern of a critical band noise for $\sigma_p^2[j] = 60$ dB

The excitation of multiple targets e_i in critical band i is defined as

$$e_i = \sum_{j=1}^{17} e(\sigma_p^2[j], i, j). \tag{6}$$

In block C of Fig. 1 the masked threshold m_i is compared to the excitation of the distortion e_i by

$$q_i = 10 \log \frac{e_i}{m_i} \qquad 1 \leq i \leq 17. \tag{7}$$

Fig. 4 depicts the ratio q_i as an example. For $e_i > m_i$ the ratio q_i is above the x-axis. If $q_i > 0$ for any i with $1 \leq i \leq 17$, we expect audible distortions in the processed signal.

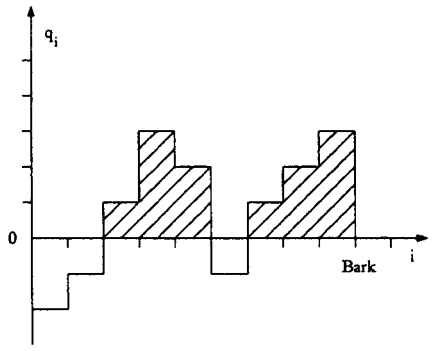

Figure 4: Example of q_i

For similar audible distortions, the hatched area in Fig. 4 monotonically increases with the degradation of the quality of the processed signal. Consequently an indication of the degradation of speech quality is obtained, which

represents the second requirement of the objective measurement.
In conclusion, the requirements for an objective quality measure are fulfilled. The quality measure $OQ(n)$ for frame n can be defined proportional to the hatched area in Fig. 4:

$$OQ(n) = \sum_{i=1}^{17} \max(0, q_i) \tag{8}$$

With N the number of frames, the mean objective quality of a processed signal is calculated as

$$\overline{OQ} = \frac{1}{N} \sum_{n=1}^{N} OQ(n) \tag{9}$$

A mean objective quality of $\overline{OQ} = 0$ indicates that the reference signal and the processed signal can not be distinguished. For larger $\overline{OQ}$ the quality of the processed signal becomes worse.
In the following section the test procedure of the objective quality measurement is described.

3. TEST PROCEDURE

For the subjective listening tests a data base was used which consists of simple, meaningful sentences of 2-3 seconds duration. They are grouped into sentences pairs. Low level background noise is added to pad out samples pairs to 8 seconds.
To select the codec candidate with the best performance, different experiments were performed. They can be distinguished by the way in which the sentences pairs were processed:

- optional IRS-prefiltering [9] to simulate the frequency characteristics of the telephone handset
- low-pass filtering and downsampling to 8 kHz
- level adjustment
- optional A-law quantization
- processing through one of the four participating codec candidates or the reference codec (GSM full rate codec) including different channel characteristics / error probabilities (EP0-EP3)
- inverse level adjustment
- upsampling to 16 kHz

The error pattern EP0-EP3 represent a realistic distribution of traffic situations for the GSM situation. They include C/I (carrier to interferer ratio) conditions assuming frequency hopping, independent Raleigh fading and static co-channel interference.
In each experiment a listener panel of 24 subjects had to rate each processed signal on a scale from 1 to 5. The average of all scores provides the so called Mean Opinion Score (MOS) [3]. This test procedure makes sure that the results of the subjective listening tests are reliable and representative.
For the objective analysis, however, informal investigations by the authors indicated that the GSM data base is not suited for an objective analysis. The pauses between two sentences pairs differ considerable in length. A candidate codec which provides bad speech quality but produces silence during pause sections is judged the better for longer pauses. In addition, each codec candidate is tested in each condition with different sentences. The audibility of a distortion at a certain time instant depends on the masked threshold which itself depends on the spoken word or phoneme at this time instance. These points introduce an inaccuracy to the objective analysis.
Therefore a special OQM data base of 32 seconds of male and 32 seconds of female German speech has been set up for the objective analysis of the GSM half rate candidate codecs. This data base was applied to all codec candidates and the reference codec.
The participating codecs were tested for a subset of conditions, i.e. four different error patterns (EP0 - EP3) were tested in combination with the IRS-prefilter, a standard PCM IIR filter as low-pass and a level adjustment of -22 dB.

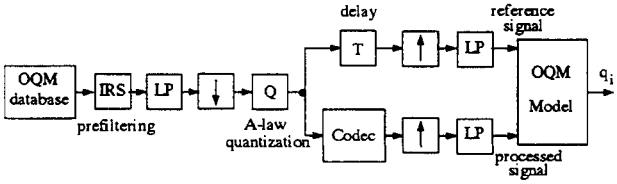

Figure 5: Test procedure for the objective analysis

The test procedure for the objective analysis is shown in Fig. 5. The lower path represents the dataflow through the codec candidate providing the files to be analysed, while in the above path the processing for the reference signal can be seen.
In the following section the results of the objective analysis are compared to the results of the subjective listening tests.

4. COMPARISON OF THE RESULTS

For the comparison of the subjective and objective measurements, the MOS-scores were extracted from the official listening test results according to the objectively tested conditions. Since each experiment was performed in two languages, the corresponding MOS values were averaged. Finally, the resulting $\overline{MOS}$ values were linearly transformed to the range 0 to 1 by

$$MOS_{comp} = \frac{\overline{MOS} - \overline{MOS}_{min}}{\overline{MOS}_{max} - \overline{MOS}_{min}} \tag{10}$$

where $\overline{MOS}_{max}$ and $\overline{MOS}_{min}$ represent the maximal and minimal value of the obtained $\overline{MOS}$ data.
Assuming the results of the subjective listening tests are reliable and representative we can compare them with the

results of the objective analysis. Consequently, the $\overline{OQ}$ values were analoguos linearly transformed according to

$$OQ_{comp} = \frac{\overline{OQ}_{max} - \overline{OQ}}{\overline{OQ}_{max} - \overline{OQ}_{min}} \qquad (11)$$

where $\overline{OQ}_{max}$ and $\overline{OQ}_{min}$ denote the maximal and minimal value of the $\overline{OQ}$ data.

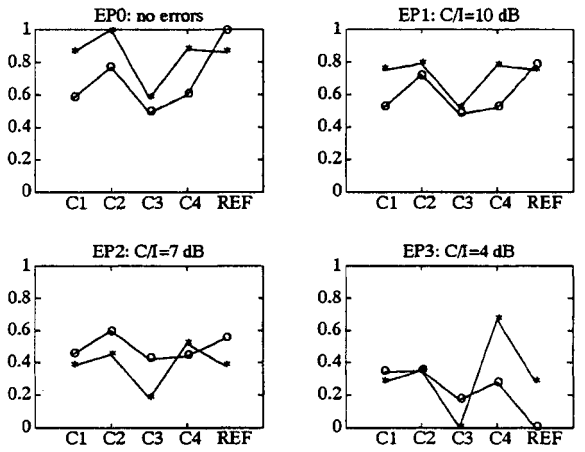

Figure 6: Comparison of the subjective test results (o) with the objective analysis (*) of 4 codecs and the GSM full rate codec as reference

Fig. 6 visualizes the comparison for the different error patterns. In agreement with the subjective listening tests, the objective quality measure detects the decreasing quality from error pattern EP0 to EP3 correctly.
Taking only the four codec candidates C1-C4 into consideration, i.e. codecs with identical brutto bitrate, both measures determine the worst performance for codec C3. In addition the rank order of the codec candidates can be predicted correctly for slightly distorted speech (EP0-EP1). However, for the remaining conditions (EP2-EP3) the correlation between both measures decreases.
For the GSM full rate codec the subjective test indicates an excellent performance for EP0 and a poor quality for EP3. This immense difference of speech quality contradicts to the objective analysis results.
In summary, especially for the error conditions EP2-EP3 the correlation between both measures is insufficient, so that a complete prediction of the subjective quality is yet not possible. According to the psychoacoustic model the OQ_{comp} value decreases monotonically with decreasing quality. However, it can not be assumed that this decrease of OQ_{comp} depends linearly on the reduction of the perceptual quality, e.g. given in term of MOS_{comp}.

5. CONCLUSIONS

An objective measure based on psychoacoustics was presented. In order to verify the performance of the objective quality measure, it was compared to the official subjective test results of the final selection test of the GSM half rate codec.

A special designed data base has been set up, which was applied to the participating codecs. Four different error pattern were tested with a special configuration of the hardware equipment.
The comparison of the official subjective listening results and the objective measure indicated a certain degree of correlation. A high correlation between both measures was observed for slightly distorted speech, a poorer correlation for larger distorted speech.
It can be concluded that

$$OQ(n) = \sum_{i=1}^{17} \max(0, 10 \log \frac{e_i}{m_i})$$

provides a suitable indication of the quality if the excitation of the distortion (e_i) is slightly above the masked threshold (m_i). For larger audible distortions $OQ(n)$ the obtained indication is insufficient. This result confirms our statement that the second requirement of an objective quality measure is difficult to meet.
We can assume that the kind of the distortion (e.g. its spectral distribution) influences the listeners' scores. Therefore, the introduction of weighting factors δ_i in form of $q_i = 10 \log(\delta_i \frac{e_i}{m_i})$ is one attempt to improve the results of the current OQM model for large distortions. However, to determine reliably the weighting factors, a larger data base is required which will be available within future subjective listening tests.

REFERENCES

[1] K. Brandenburg, T. Sporer, "NMR and Masking Flag: Evaluation of Quality Using Perceptual Criteria", *Proc. of 11th International AES-Conference*, Portland, Oregon, 1992

[2] J.G. Beerends, J.A. Stemerdink, "Measuring the quality of audio devices", *90th AES-Convention*, Paris 1991, Preprint 3070

[3] ETSI/TM/TM5/TCH-HS Expert Group Traffic Channel Half Size, "GSM Half Rate Selection Test Plan", TD 92/14, April 1992

[4] K. Brandenburg, et al., "The ISO/MPEG-Audio Codec: A generic standard for coding of high quality digital audio", *92nd AES-Convention*, Vienna, 1992

[5] R.N.J. Veldhuis, "Bit Rates in Audio Source Coding", *IEEE Journal on Selected Areas in Communications*, Vol. 10, No. 1, Jan. 1992

[6] E. Zwicker, H. Fastl, "Psychoacoustics", Springer Verlag, Berlin, 1990

[7] J.D. Johnston, "Estimation of Perceptual Entropy Using Noise Masking Criteria", *Proceedings of the ICASSP*, 1988

[8] Boff, Kaufmann, Thomas "Handbook of Perception", Vol. 1, pp. 14-36-14-38, 1986

[9] ETSI/TM/TM5/TCH-HS Expert Group Traffic Channel Half Size, "Proposal of a digital IRS filter specification for sampling rates at 8 kHz and 16 kHz", TD 90/29, July 1990