# IMPLEMENTATION OF A PSYCHOACOUSTICAL PREPROCESSING MODEL FOR SOUND QUALITY MEASUREMENT

Martin Hansen & Birger Kollmeier
Email: martin@medi.physik.uni-oldenburg.de
AG Medizinische Physik, Carl von Ossietzky-Universität Oldenburg
D-26111 Oldenburg
GERMANY

## ABSTRACT

This study investigates the applicability of a psychoacoustical preprocessing model for measuring the speech quality of low-bit-rate codecs. The method for deriving an objective measure of the sound quality is to transform the input and output signal of a speech coding device into a so-called "internal representation" of the sound. Differences between the internal representations of input and output signal, calculated as the correlation coefficient between them, correspond to a decreased quality of the output signal. The preprocessing model consists of a critical-band filterbank which simulates the filtering by the basilar membrane, followed by a nonlinear adaptation circuit modeling amplitude compression and temporal masking effects, and a further filtering stage which analyzes fluctuations of the temporal envelope of the sound signal. This signal processing seems to be a successful "universal" model to simulate a variety of psychoacoustical data as well as the perception of speech sound at the same time.

## 1 INTRODUCTION

Methods for objective ("instrumental") speech quality measurement have gained increasing importance for the development and evaluation of new low-bit-rate speech coding algorithms and devices, since subjective listening tests (as carried out for standardization tasks) are very time consuming and need a substantial effort in order to yield a good reliability and reproducibility. Recently, psychoacoustical knowledge about features of the peripheral signal processing in the auditory system is being included in the methods for speech quality measurement [1, 2, 3, 4, 5]. At the same time, a successful speech quality measure gives important new insights into the mechanisms of speech perception under (partly) adverse conditions and can answer the question, which features of a speech signal are perceptually relevant.

The aim of this work is to investigate the applicability of an existing psychoacoustic signal processing model for the needs of objective speech quality measurement. In our work group an elaborated functional model for the signal preprocessing has been developed which originated from a former group at University of Göttingen. This model has been successfully applied to conditions of simultaneous and non-simultaneous masking in a wide variety of psychoacoustic experiments [6, 7]. The signals applied in these experiments consist of well-defined components like sinusoidals, harmonic tone complexes and (filtered) noise. The question arises, if the model for the detection and discrimination of these signals is sufficient to model the much more complex problem of (specific) tasks of speech perception.

In this study the existing signal preprocessing model was applied in a way to measure the speech quality of low-bit-rate speech coding-encoding devices ("codecs") such as used in mobile telephony. These codecs produce a speech signal that is fully intelligible and allows for almost normal speaker identification, compared to standard telephony, but exhibit a clearly reduced speech quality due to its highly non-linear and/or time-variant signal transformation algorithm.

The goal in objective speech quality measurement is to quantify the quality degradation of a speech sample relatively to an undegraded reference situation. The application of psychoacoustical preprocessing models is motivated by the assumption that subjects are able to judge the quality of a test speech signal by comparing a kind of "internal perceptual representation" of the test sound with that of a reference sound. This representation is thought of as the information that is accessible to higher neuronal stages of perception. It should contain the perceptually relevant features of the incoming sound. Differences in this "internal representation" of input and output signal are expected to correspond to perceivable differences of the two signals and thus indicate a decreased speech quality of the output signal. In the case of subjective listening tests for low-bit-rate speech codecs, the references are not presented to the subjects. However, it can be assumed that subjects can make use of an internally available (approximate) template of the reference situation. Therefore, in the objective method the reference signal is used as well to represent the the a-priori knowledge of the listener about the undegraded speech.

# 2 METHOD

## 2.1 Global adjustment

The input and output signal of the device under test are globally time-aligned and linearly scaled to have the same overall rms-value prior to the computing of the "internal representations" of each of both signals. The preprocessing model is then applied to transform the input and output signal to their so-called internal representations.

## 2.2 Basilar-membrane filtering

The first stage consists of a gammatone band-pass filterbank [8]. The filterbank is optimized to match the shape of the impulse response of the auditory filters on the basilar membrane. The signal is split up into 19 band-pass signals with center frequencies from 300 to 3460 Hz, equally distributed with 1 filter per ERB (equivalent rectangular bandwidth).

## 2.3 Nonlinear compression and adaptation

Subsequent to the filterbank analysis, a nonlinear amplitude compression is applied to each channel. This stage models the approximately logarithmic level compression for stationary input.
Three different processing algorithms were investigated.

### 2.3.1 Perception model

A block-diagram of the signal processing of the perception model is shown in Fig. 1. For each channel the transduction characteristic of the inner hair cells is modeled by a halfway-rectification followed by a low pass filter at 1 kHz. For high frequencies this resembles an envelope extraction, while for lower frequencies additionally the phase information of the input signal is preserved. At this stage adaptation loops, developed by Püschel [9] to model forward masking, are applied to each channel to model nonlinear amplitude compression and temporal masking effects. These adaptation loops are build up of five similar division elements, in which the fedback and low-pass filtered output of each element serves as the divisor of that element. Stationary signals are compressed in their amplitude according to $y = \sqrt[32]{x} \approx \log x$, as each element gives the square root of its input if in steady state. Fluctuations in the input signal are first passed through this system linearly, but change the charging state of the low pass filters at the same time. Thus the compressive characteristic is changed adaptively depending on the signal's amplitude at previous times. The five time constants are spaced linearly between the fixed two values of $\tau_1$=5 ms and $\tau_5$=500 ms. A low-pass filter at 8 Hz limits the temporal resolution in the adapted compressed output. Finally each channel is block-averaged in frames of 20 ms duration. The result of this computation is called the internal representation of the signal.

Besides the "perception model" described in section 2.3.1 two alternative, simpler algorithms for amplitude compression were investigated. These algorithm differ fro the former only in the stage between basilar-membrane filtering and the final block-averaging in 20 ms frames.

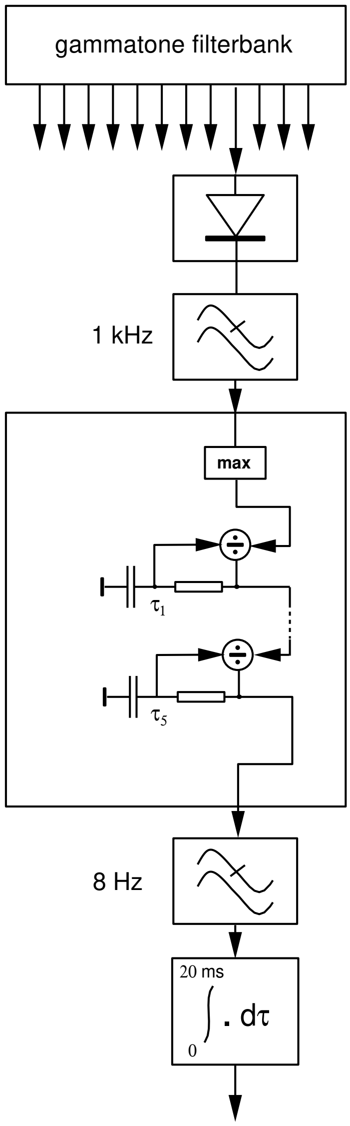


Figure 1: Signal processing scheme of the perception model. Transformation of the input sound signal into its internal representation.

### 2.3.2 Adapted bark spectral distortion

In each filterbank channel the energy is calculated in 20 ms frames. The energy is compressed by a power law $y = x^{0.23}$, according to Zwicker's functional relation between loudness and intensity of a stationary sound.

### 2.3.3 Instantaneous logarithmic compression

In each channel, the samples are halfway-rectified, subjected to an instantaneous logarithmic compression, and low-pass filtered at 8 Hz. Each channel is block-averaged in 20 ms frames.

## 2.4 Calculation of the quality measure

The internal representation is calculated in the same way for the preadjusted input and output signal of a codec device under test. The internal representations are frequency weighted according to the ISO 226 isophone for 40 phon at the 19 center frequencies of the basilar-membrane filterbank. From these two weighted representations, the objective speech quality measure $q$ is then calculated as the correlation coefficient between all corresponding samples of the two-dimensional time-frequency representations: Let $X(i,j)$ and $Y(i,j)$ be the representations at times $t = i \cdot \Delta t, i = 1 \dots N$ and frequencies $f = j \cdot \Delta f$, $j = 1 \dots M$. The speech quality measure $q$ is then defined as

$$q = \frac{\sum_{i,j=1}^{N,M} (X(i,j) - \overline{X}) \cdot (Y(i,j) - \overline{Y})}{\sqrt{\sum_{i,j} (X(i,j) - \overline{X})^2} \sqrt{\sum_{i,j} (Y(i,j) - \overline{Y})^2}},$$

where $\overline{X}$ and $\overline{Y}$ denote the means of $X(i,j)$ and $Y(i,j)$.

# 3 RESULTS

## 3.1 Test material

For each of the three preprocessing models the method was applied to the speech material of three different tests of comparable number and lengths of the included sentences. All the subjective listening tests were carried out at and the speech material was made available by the Research Center of Deutsche Telekom AG, Berlin. The first test was the "ETSI Halfrate Selection Test" in German language. Another two tests were internal tests [10] for the quality evaluation of multiply cascaded ADPCM-codecs, one test ("C 1") with simulated, the other ("C 2")with real net-connections. In all three tests, always four sentences of different speakers were used to assess the overall subjective quality of each of different codec conditions on a scale with five absolute categories. The subjective speech quality is calculated as the mean of the judgments of 24 subjects (mean opinion score "MOS"). For the objective measurement, these four sentences were concatenated to one continuous signal and the mean of the MOS was taken.

The results of the objective speech quality calculation for the preprocessing by the perception model ("PM") are shown in Fig. 2. MOS is plotted versus the objective measure $q$. The numbers and characters refer to the numbers of the codecs or to cascading-conditions. The special character m refers to a MNRU-reference codec system.

The quality of a speech quality prediction by the objective measure $q$ can be estimated by the rank-correlation coefficient between $q$ and the subjective speech quality, and the standard deviation of a fitted function for the data. These values are given in the upper left corner of each panel (rs, SD). The results of the speech quality prediction of the two alternative preprocessing models, bark spectral distortion ("BSD") and instantaneous logarithmic compression

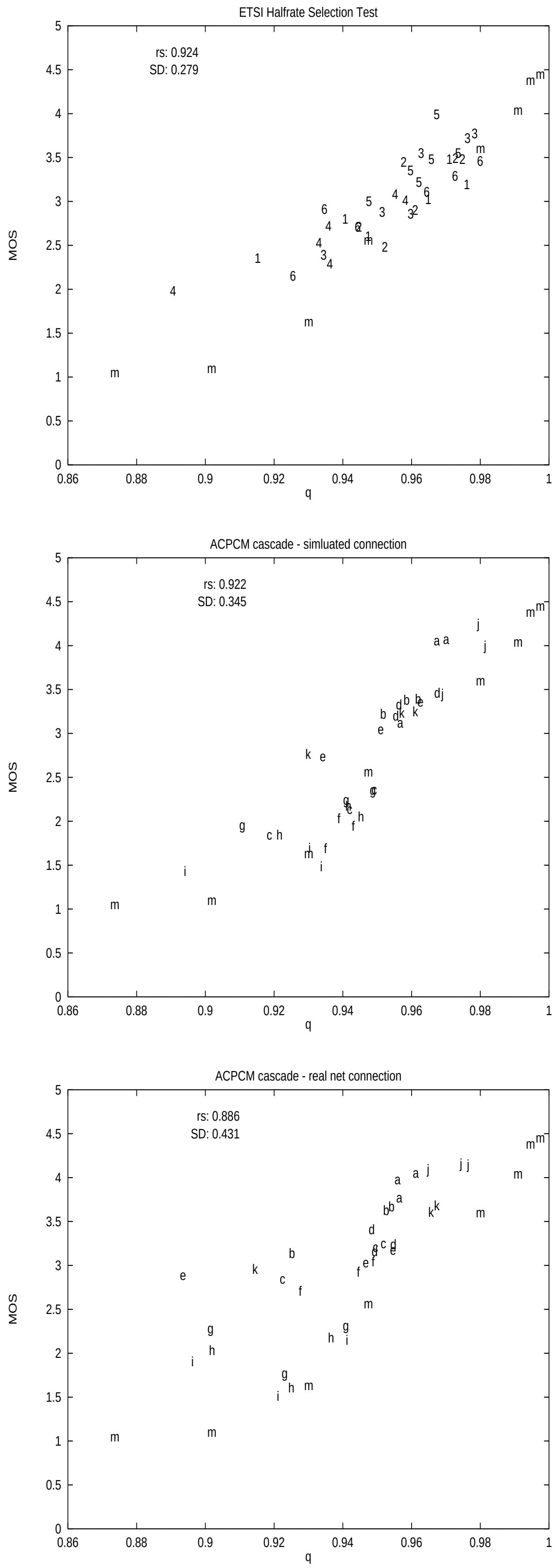


Figure 2: Results of the objective speech quality measurement with perception model preprocessing: Subjective quality (MOS) versus objective measure $q_{PM}$ for the ETSI Halfrate Selection Test and two cascaded-ADPCM tests.

("LOG"), are are not shown graphically but are summarized in table 1.

| Test | PM | | BSD | | LOG | |
|---|---|---|---|---|---|---|
| | rs | SD | rs | SD | rs | SD |
| ETSI | 0.924 | 0.279 | 0.842 | 0.370 | 0.866 | 0.463 |
| C 1 | 0.944 | 0.366 | 0.851 | 0.547 | 0.937 | 0.609 |
| C 2 | 0.898 | 0.501 | 0.823 | 0.560 | 0.869 | 0.576 |

Table 1: Rank-correlation coefficient (rs) and standard deviation (SD) of a second order polynomial fit between subjective quality and objective measure $q$.

A comparison across the different preprocessing models and tests shows the following results: For all three test $q_{PM}$ yields the highest rank-correlation coefficient and the lowest standard deviation compared to $q_{LOG}$ and $q_{BSD}$. For the ETSI and the simulated cascading test the $q_{PM}-$ MOS data can be fitted by a monotonic function with only small deviations. In particular, no clusters occur in the $q_{PM}-$ MOS data. This means that the individual signal degradations introduced by the different types of codecs are transformed in a perceptually "correct" way into the internal representations. For the real-net cascading test, some larger deviation and a clustering of the $q_{PM}-$ MOS data occurs, at low values of $q_{PM}$, leading to a lower rank-correlation coefficient and higher standard deviation. Some signal degradation features of the real-net environment may not yet be sufficiently described in the internal representation of these signals. For $q_{LOG}$ and $q_{BSD}$ the graphical analysis shows similar results as in the panels of Fig. 2, but exhibits larger deviations and stronger clustering of the MRNU-reference codec system compared to the other codecs under test. However, it is surprising, that for the simulated-connection test also the relatively simply $q_{LOG}$ is a good predictor of the subjective speech quality (rs=0.937), although with higher standard deviation compared to $q_{PM}$.

The preprocessing models differ in the way the amplitude compression is performed. The realistic dynamic compression, accounted for by the adaptation loops, may explain why the perception model preprocessing performs best. This compressive characteristic is appropriate *both* for stationary parts of the signal where the input $x$ is transformed approximately logarithmically, *and* for sections of the signal exhibiting temporal masking effects (e. g. backward-, simultaneous-, forward masking and gaps). This conclusion is supported by the speech recognition experiments of Tchorz [11]. The same perception model preprocessing as described in this paper was applied as a feature extraction method for speech recognizing algorithms. In conditions with noisy background, the speech recognition rates were more robust than standard LPC- or RASTA-PLP-coefficients .

## Acknowledgment

This work has been supported by the Research Center of Deutsche Telekom AG.

## References

[1] S. Wang, A. Sekey, and A. Gersho. "Auditory Distortion Measure for Speech Coding". In *IEEE Proc. Int. Conf. Acoust., Speech Signal Processing*, pages 493–496, 1991.

[2] J. G. Beerends and J. A. Stemerdink. "A Perceptual Speech Quality Measure based on a Psychoacoustic Sound Perception". *J. Audio Eng. Soc.*, 42 (3):115–123, 1994.

[3] J. Berger and A. Merkel. "Psychoakustisch motivierte Einzelmaße als Ansatz zur objektiven Qualitätsbestimmung von ausgewählten Sprachcodiersystemen". In *Elektronische Sprachsignalverarbeitung, Proceedings*, TU-Berlin, 1994.

[4] M. Hansen and B. Kollmeier. "Objektive Sprachqualitätsvorhersage mittels einer gehörorientierten Vorverarbeitung". In *Fortschritte der Akustik DAGA '96*. Dega, 1996.

[5] M. Hansen and B. Kollmeier. "Prediction of Speech Quality based on Psychoacoustical Preprocessing Measures". In *Workshop on Quality Assessment in Speech, Audio and Image Communication, Proceedings*, pages 7–12, Darmstadt, 1996. ITG/EURASIP.

[6] T. Dau, D. Püschel, and A. Kohlrausch. "A quantitative modell of the "effective" signal processing in the auditory system: I. Model structure". *J. Acoust. Soc. Am.*, 1996. in press.

[7] T. Dau, D. Püschel, and A. Kohlrausch. "A quantitative modell of the "effective" signal processing in the auditory system: II. Simulations and measurements". *J. Acoust. Soc. Am.*, 1996. in press.

[8] R. Patterson, Nimmo-Smith I. J. Holdsworth, and P. Rice. "An efficient auditory filterbank based on the gammatone function". In *Meeting of the IOC Speech Group on Auditory Modelling at RSRE*, pages 14–15, 1987.

[9] D. Püschel. "Prinzipien der zeitlichen Analyse beim Hören". PhD thesis, Universität Göttingen, 1988.

[10] J. Berger. "Ein Ansatz zur instrumentellen Sprachqualitätsabschätzung in Telefonverbindungen im Festnetz der Deutschen Telekom". In *Workshop on Quality Assessment in Speech, Audio and Image Communication, Proceedings*, pages 17–24, Darmstadt, 1996. ITG/EURASIP.

[11] J. Tchorz, M. Wesselkamp, and B. Kollmei er. "Gehörgerechte Merkmalsextraktion zur robusten Spracherkennung in Störgeräuschen". In *Fortschritte der Akustik DAGA '96*, page im Druck. Dega, 1996.