USING A QUANTITATIVE PSYCHOACOUSTICAL SIGNAL REPRESENTATION FOR OBJECTIVE SPEECH QUALITY MEASUREMENT

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

ABSTRACT

This paper describes the application of a quantitative psychoacoustical signal preprocessing model for objective speech quality measurement. The preprocessing is applied to transform the original and the distorted speech signal to an internal representation which is thought of as the information that is accessible to higher neural stages of perception. From a comparison of these internal representations a quality measure can be derived that shows a high correlation to the subjective MOS data of various test data bases. The inherent parameters of the preprocessing model were derived directly from psychoacoustical data independent of the present study. The detection thresholds of codec-like distortions obtained in a psychoacoustical experiment could also be predicted by the model. This indicates that the internal representation contains the relevant information for detecting perceivable differences. It provides evidence for a direct relation between speech quality and detectability of a distortion.

1. INTRODUCTION

The goal in speech quality measurement is to quantify the quality degradation of a speech sample relatively to an undegraded reference situation in order to predict the subjective quality rating of a listener panel. Objective, i.e. non-auditive, instrumental methods for speech quality measurement have been of large interest for a long time already due the substantial effort needed for reliable and reproducible subjective listening tests.

In this study, the distorted speech signals were generated by low-bit-rate speech coding-decoding 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 their highly nonlinear and/or time-variant algorithm. In listening experiments, carried out by the research center of Deutsche Telekom, the speech quality was subjectively rated by test subjects.

In objective speech quality measurement, 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 [1, 2, 3, 4, 5]. This representation is thought of as the information that is accessible to higher neural 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 to indicate a decreased speech quality of the output signal.

2. PSYCHOACOUSTICAL PREPROCESSING MODEL

An elaborated functional model of the auditory processing has been successfully applied by Dau [6] to simulate conditions of simultaneous and non-simultaneous masking in a wide range of psychoacoustical experiments. As a first stage, a gammatone-filter bank simulates the filtering by the basilar membrane [7]. It splits the signal into 19 critical bands with center frequencies from 350 to 3500 Hz. Each channel is halfwave rectified and low-pass filtered at 1 kHz to model the hair cell transduction characteristic. The adaptation loops [8], modeling temporal masking effects, calculate a nonlinearly compressed and adapted envelope of the signal. A final low-pass filter at 8 Hz analyzes the temporal resolution of the envelope signal.

3. APPLICATION OF THE MODEL TO SPEECH QUALITY MEASUREMENT

The above model was applied to speech quality measurement in the way depicted in Fig. 1.

In subjective listening tests, typically sentences of different speakers are coded by the same codec-condition and are rated individually by the subjects. For the objective method, these sentences of different speakers were concatenated and their average mean opinion score (MOS) calculated in order to reduce the variability of the MOS due to the different voices of the speakers. The original and the distorted signal are then aligned with respect to overall delay and overall RMS. Both signal are then transformed to their internal representations.

For the purpose of speech quality measurement, each channel output is time-averaged in frames of 20 ms with 50% overlap. A band weighting (see Fig.2) is applied by individual gain factors for each channel. This weighting characteristic accounts for the relative perceptual importance of different band for speech quality. It is a slight modification of the 40-Phon Iso-loudness contours. The final objective speech quality measure q is calculated as the overall correla-

tion coefficient between the weighted representations of the original and the distorted signal. A value of $q = 1.0$ reflects identical representations. The quality in terms of MOS is expected to be a monotonically increasing function of q.

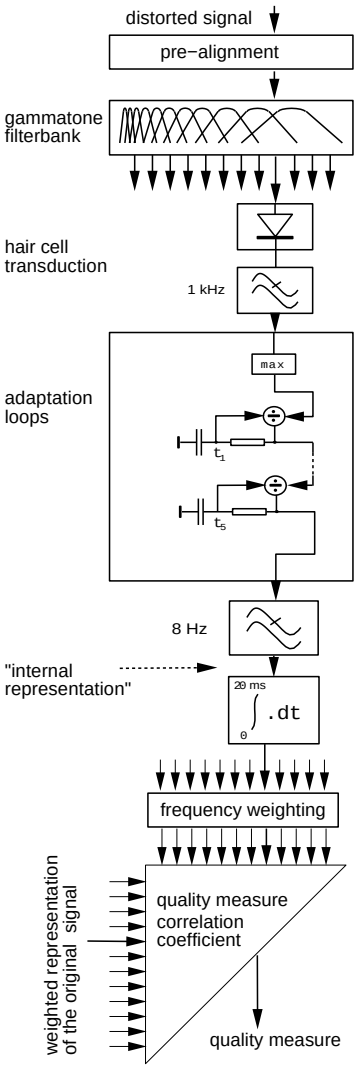

Figure 1. Signal processing scheme of the perception model. Transformation of the input sound signal into its internal representation according to [6], and calculation of the speech quality measure.

4. RESULTS WITH THE OBJECTIVE SPEECH QUALITY MEASURE

The described method was applied to the German language test material of the ETSI GSM-halfrate selection test, the ITU-8kbit test and two internal tests with differently cascaded ADPCM, the latter performed by Deutsche Telekom. The results of the objective speech quality measurement are shown in Fig. 3. The subjectively rated quality in terms of the MOS is plotted versus the objective quality measure. The numbers and characters represent the different codec

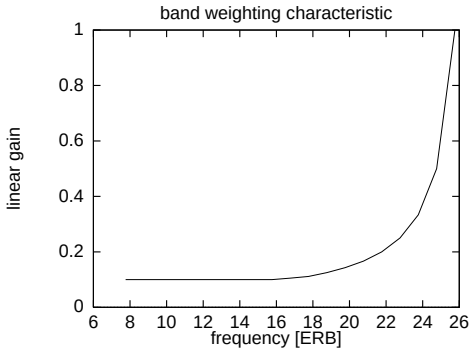

Figure 2. Band weighting applied to the internal representation before calculation of q, as a function of the critical bandwidth ERB.

conditions within the corresponding test. The correlation coefficient and the rank correlation coefficient between subjective and objective data and the standard deviation of a first order fit function are given in the upper left of each panel.

For all four tests, the data can be fitted by a monotonic function with only small deviations. In particular, no clusters occur. This means that the individual signal degradation introduced by the different types of codecs seem to be transformed in a perceptually "correct" way at the stage of the internal representations. In all four tests, the correlation coefficients are at high levels above 0.9 and the standard deviations are smaller than the average standard deviations of the interindividual variations of the test subjects.

Additionally it can be observed, that not only a high correlation coefficient within the data of one test data base is acchieved but also that a certain range of q-values corresponds to the same range of MOS across the different test data bases, which indicates that q is a good general objective predictor of the speech quality.

The results were compared with alternative methods of objective speech quality measurement from the literature [1, 2, 3]. A comparison showed (data not presented here), that the presented method performed best in most cases. Only the PSQM [2], in the ITU 8kbit test data base for Dutch language, yielded similar results, in terms of the correlation coefficient between subjective and objective data.

In a post-hoc analysis of the model parameters employed here it was found that all variations of the model parameters led to a deterioration in speech quality estimates. This indicates that the same parameterset which was optimized in typical psychoacoustical simulations can also serve as an optimal setup for speech quality measurement.

5. DETECTION THRESHOLD OF BAND-SPECIFIC DISTORTIONS

In the development of the objective speech quality measure q it was assumed that perceivable differences between two test signals correspond to differences in the internal representations of the signals.

In a further experiment it was investigated how the objective speech quality measure q described above could predict the *detectability* of a frequency-dependent distortion intro-

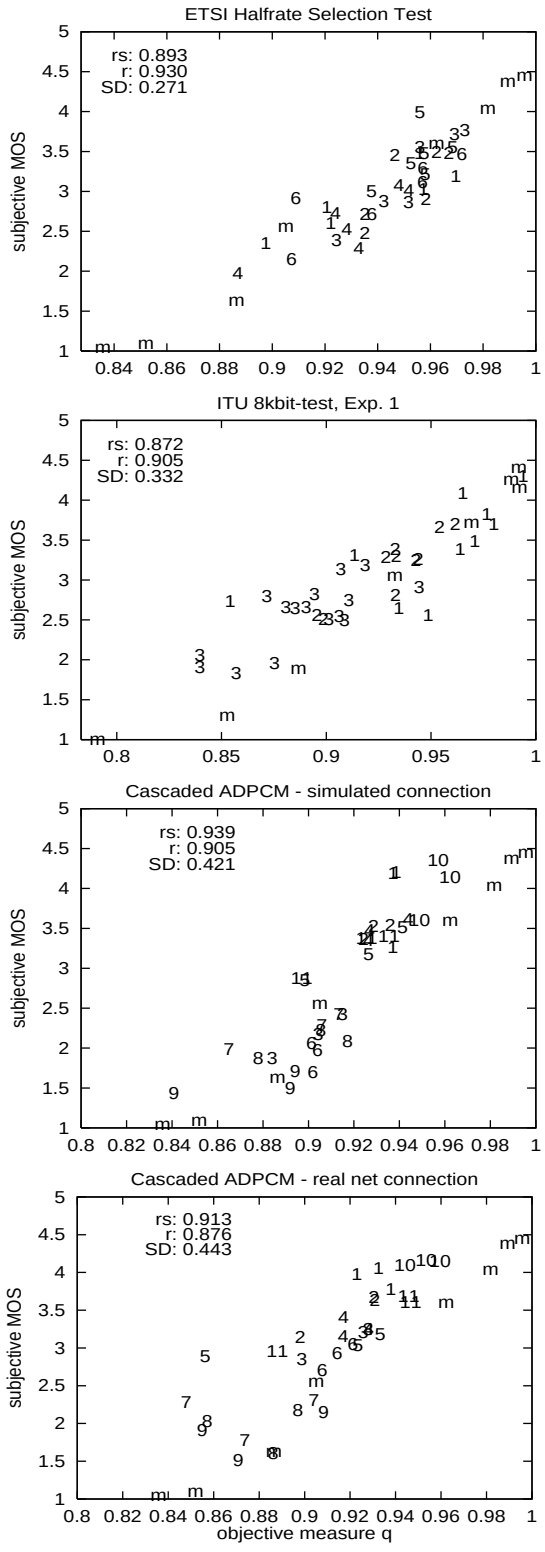

Figure 3. Results of the objective speech quality measurement with auditory preprocessing model: Subjective quality (MOS) versus objective measure q for the ETSI Hafrate Selection Test, ITU 8kbit test, and two cascaded ADPCM test.

duced into the speech signal.

Two reference sentences of 2 s duration, uttered by a male speaker, were chosen from the ETSI data base material. To introduce a band-specific, codec-like distortion, the reference speech signal was band-pass filtered at a given center frequency f_c, and subsequently modulated by a Modulated Noise Reference Unit (MNRU) [9], with a modulation depth of $-Q$ dB [10]. The bandwidth of the filters was approximately one critical band. The original signal was notch filtered, with corresponding center frequency and bandwidth, and added to the modulated signal (Fig. 4). The distortion is broadband, carrying information of a specific narrow band of the input speech signal.

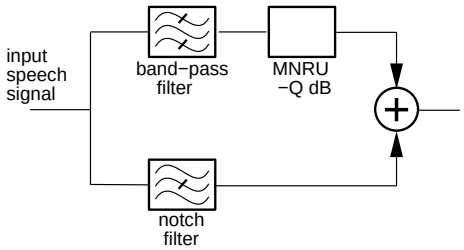

Figure 4. Signal generation for the band-specific modulated distortion.

In a psychoacoustic experiment, the reference signal and the manipulated sentence were presented to the subjects in a 2I-2AFC 1-up-2-down paradigm. Detection thresholds of the noise-modulated test signal were measured in terms of the modulation depth as a function of the center frequency. Two subjects participated in this experiment. The center frequencies were set to 300, 500 and 700 Hz, and, with equidistant spacing of 500 Hz, from 1 kHz upto 3.5 kHz.

The left and right panel of Fig.5 show the results for the two different sentences.

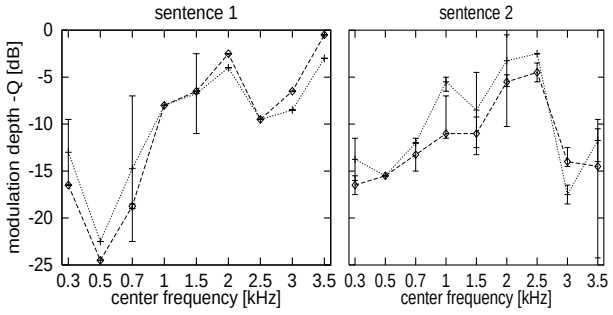

Figure 5. Detection threshold for subjects MH (dashed) and RK (dotted) for the band-pass filtered modulated signal in the presence of notched original signal.

The objective speech quality measure q was calculated for the same stimuli that were generated in the psychoacoustical experiment, with a flat band weighting instead of the one shown in Fig.2. The unmodulated signal was taken as the reference. For the two different sentences, the modulation depth parameter Q of the MNRU leading to a constant value of the objective speech quality q was determined as a

function of f_c. The iso-q-value contours for three different levels of q are shown in Fig.6 for sentence 1 and sentence 2 (upper and lower panel). From this figure it is obvious that the iso-q-value contour for q=0.999 represents the measured threshold data very well for both sentences and across all frequencies.

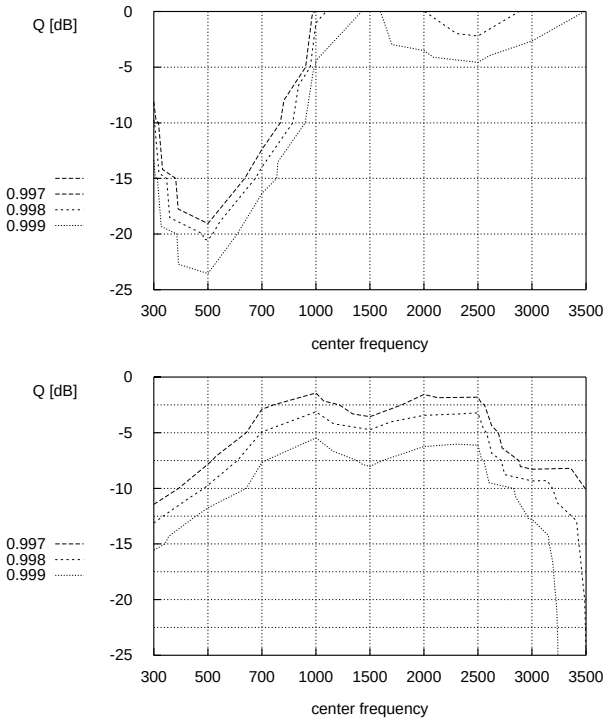

Figure 6. Iso-q-value contour for the stimuli used in the experiment. The modulation depth Q leading to a constant q-value of 0.999, 0.998, and 0.997 is shown in the graph.

6. CONCLUSION

The objective-subjective speech quality data can be fitted by an monotonic function with only small deviations. A high correlation coefficient is acchieved and, in particular, no clusters of different codec types occur. 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.

From the psychoacoustical experiment, it can be concluded that the objective measure q also provides a good prediction of the *detection threshold* of introduced band-specific distortion. This finding supports the basic model assumption, that the internal representations in fact contains the relevant information for detection of perceivable differences.

The realistic dynamic compression, accounted for by the adaptation loops, may explain why the perception model preprocessing performs best compared to alternative methods. The 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 which is always found in speech signals.

In sum, the model shows to be a unifying approach for a variety of psychoacoustical and speech processing tasks.

ACKNOWLEDGEMENT

The authors wish to thank Reinier Kortekaas for fruitful discussions and suggestions concerning the detection experiment.

The subjective speech quality tests were performed at and the speech data base material was made available by the Deutsche Telekom AG. Parts of the study have been financially supported by the 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. "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.

[5] M. Hansen and B. Kollmeier. "Implementation of a psychoacoustical preprocessing model for sound quality measurement". In *ESCA Tutorial and Workshop on The Auditory Basis of Speech Perception*, pages 79–82, Keele, 1996.

[6] T. Dau, D. Püschel, and A. Kohlrausch. "A quantitative model of the 'effective' signal processing in the auditory system: I. Model structure". *J. Acoust. Soc. Am.*, 99:3615–3622, 1996.

[7] R. Patterson, Nimmo-Smith I. J. Holdsworth, and P. Rice. "An efficient auditory filterbank based on the gammatone function". In *Appendix B of SVOS Final Report: The auditory Filterbank*. APU report 2341, 1987.

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

[9] CCITT Blue Book Vol. V Rec. P.81. "Modulated Noise Reference Unit (MNRU)", 1989.

[10] M. Hansen and B. Kollmeier. "On the relative importance of individual critical bands for the perception of speech quality". In *Contributions to Psychological Acoustics*, BIS Universität Oldenburg, 1996. Summer school of the Oldenburg Graduate collegue psychoacoustics. In press.