APPLICATION OF MEDDIS' INNER HAIR-CELL MODEL TO THE PREDICTION OF SUBJECTIVE SPEECH-QUALITY

Markus Hauenstein

Institute for Network & System Theory, University of Kiel,
Technical Department, Kaiserstrasse 2, 24143 Kiel, Germany,
Tel: +49 431 77572-405; fax: +49 431 77572-403; e-mail: mh@techfak.uni-kiel.de

ABSTRACT

This paper demonstrates how an instrumental speech-quality measure based on the comparison of auditory-nerve firing-patterns can be constructed. Four available subjective tests prove that the mean opinion scores (MOS) estimated by the objective measure are in good agreement with the subjectively obtained results.

1. INTRODUCTION

The commonly used subjective way of assessing the speech quality of a speech codec consists of asking some people to listen to speech samples that were processed by the codec and to grade them. Usually a scale from 1 (poor quality) to 5 (good quality) is used. The different scores are collected and a mean opinion score (MOS) is calculated for the codec. Objective methods, however, aim to replace these time-consuming and expensive subjective tests by an instrumental measure, i.e. a computer program.

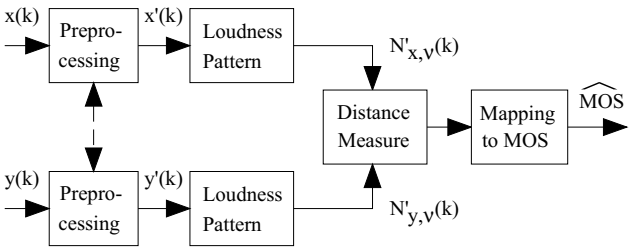


Figure 1: Basic structure of an instrumental speech-quality measure comparing specific-loudness patterns as internal signal representations.

Figure 1 shows the basic structure of a 'state-of-the-art' objective speech-quality measure which is comparing specific-loudness patterns of the codec input signal $x(k)$ and the output signal $y(k)$. These representations of the speech signals in a 3-dimensional space (specific-loudness versus warped frequency versus time) are closer related to the human speech perception than the corresponding time signals [3]. Thus audible degradations should be more clearly expressed in this perception domain than in the time or frequency domain.

The basic idea of such a measure was first presented in [4] and afterwards picked up by other researchers. Recently, ITU-T has standardized a comparable measure called PSQM for the objective quality measurement of coded speech [5, 6]. Although all these measures correlate in some cases quite good with subjectively obtained scores, unfortunately quite often the correlations between subjective and instrumental results remain unsatisfactory, and the search for better instrumental measures continues.

The use of specific-loudness patterns as an internal signal representation is motivated by results delivered by psychoacoustics. This field of research tries to link physically measurable parameters of sound waves with the human auditory sensations that they create. Psychoacoustics provide black-box models, and the anatomical and physiological mechanisms that mediate the sensations are therefore of minor interest in psychoacoustics. Specific loudness is a hypothetical quantity developed primarily for the estimation of perceived loudness (through integration of specific loudness), and it may be asked if small signal degradations can be accurately detected in this way especially when the signal statistics are rapidly changing as in speech. In contrast to specific loudness, nerve spikes can be measured by physical means. The firing probabilities of the auditory nerve fibers constitute thus a physically meaningful signal representation well adjusted to the information flow towards the human brain, and provide therefore an alternative internal representation of sound events possibly leading to better instrumental measures. In practice, the creation of nerve spikes can be modeled with inner hair-cell models. Figure 4 shows an experimental instrumental measure comparing auditory-nerve firing-patterns instead of specific-loudness patterns. The de-

tails of this measure are descriped in the next sections.

2. PREPROCESSING

Before the firing probabilities are calculated, the undistorted codec input signal $x(k)$ and the coded/decoded and therefore degraded codec output signal $y(k)$ have to pass through several preprocessing stages: The gain and the delay of the codec are measured and compensated, then a voice-activity detection (similar to the algorithm used in the GSM-networks) eliminates the speech pauses which should have no influence on the perceived speech quality. The reduced bandwidth of the telephone channel is afterwards modeled by an FIR-bandpass filter, and a second FIR-filter approximates the frequency response of the average telephone handset modeling thus the telephone situation in the subjective test.

3. FIRING PATTERNS

The firing patterns of the preprocessed signals $x'(k)$ and $y'(k)$ can now be calculated. Before sound waves are analyzed by the hair-cells in the inner ear (cochlea), they have to pass the outer and the middle ear. This transfer can be modeled by a linear time-invariant filter, and a 4th-order IIR-filter is sufficient to approximate the outer- to inner-ear transfer function (see Figure 2).

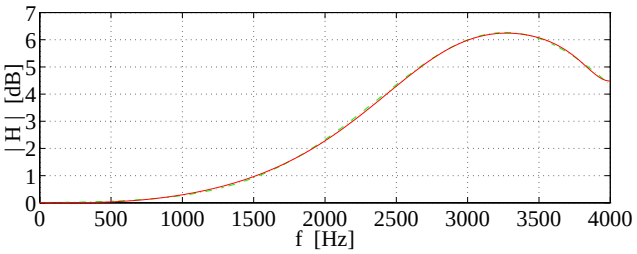


Figure 2: Outer-inner ear frequency-response. Solid: 4th-order IIR-filter, dashed: analytical formula close to measured data.

The next step (after an oversampling which is required because the hair-cell models do not work properly if the sampling rate is too low) consists of modeling the peripheral auditory filtering performed by the basilar membrane in the inner ear. This task can be accomplished with the popular gammatone filterbank [2]. On the basilar membrane we find a row of inner hair-cells which constitute the actual auditory receptors. Their hairs are sheared by the basilar membrane motion which leads - through physiological processes - to the generation of electrical impulses in the efferent nerve fibers connected to the hair-cell bodies. Thus, acoustical information is transcoded into a train of nerve spikes that is passed on to the brain. Several models of this mechanical to neural transduction have been described in literature (see e.g. [9]), and a popular and well documented one is the inner hair-cell model developed by Meddis [7]. Each output signal of the gammatone filterbank is therefore driving an inner hair-cell model as it was proposed in [8], which leads to a firing probability versus basilar membrane place versus time representation of the analysed speech signal.

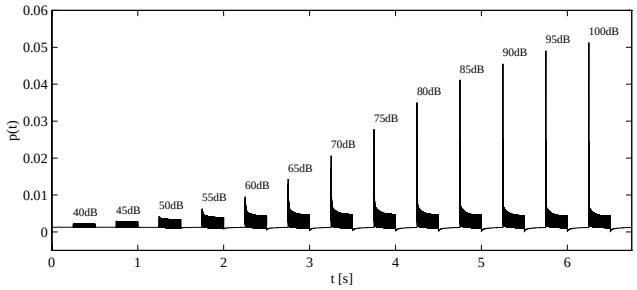


Figure 3: Reaction of Meddis' hair-cell model to sinusoidal excitation: 13 sinusoids having a duration of 250 ms each and separated by pauses of 250 ms were used as input to the model.

The firing probabilities supplied by the hair-cell models show a pronounced temporal fine structure. A direct comparison of the probabilities of the input and the output signal whould hence lead to big distances due to small but inaudible phase differences. The temporal fine structure is therefore smeared through lowpass-filtering. This leads to the final internal representations $p_{x\nu}(\kappa)$ of the original and $p_{y\nu}(\kappa)$ of the processed speech signal which must be compared using an adequate distance measure.

4. DISTANCE MEASURE

The distance measure models the information processing in the human brain. This must be regarded as a rough simplification, but since the complicated auditory processing stages in the human brain are far from being finally explored and a satisfying model will not be available in the short or even long term, there is no other choice than trying out some sensible distance measures (at most, some simple effects can be taken into account) and using the one that leads to the highest correlations with subjective scores. The distance measure should not have too much parameters, otherwise we would risk to interpolate between instrumental

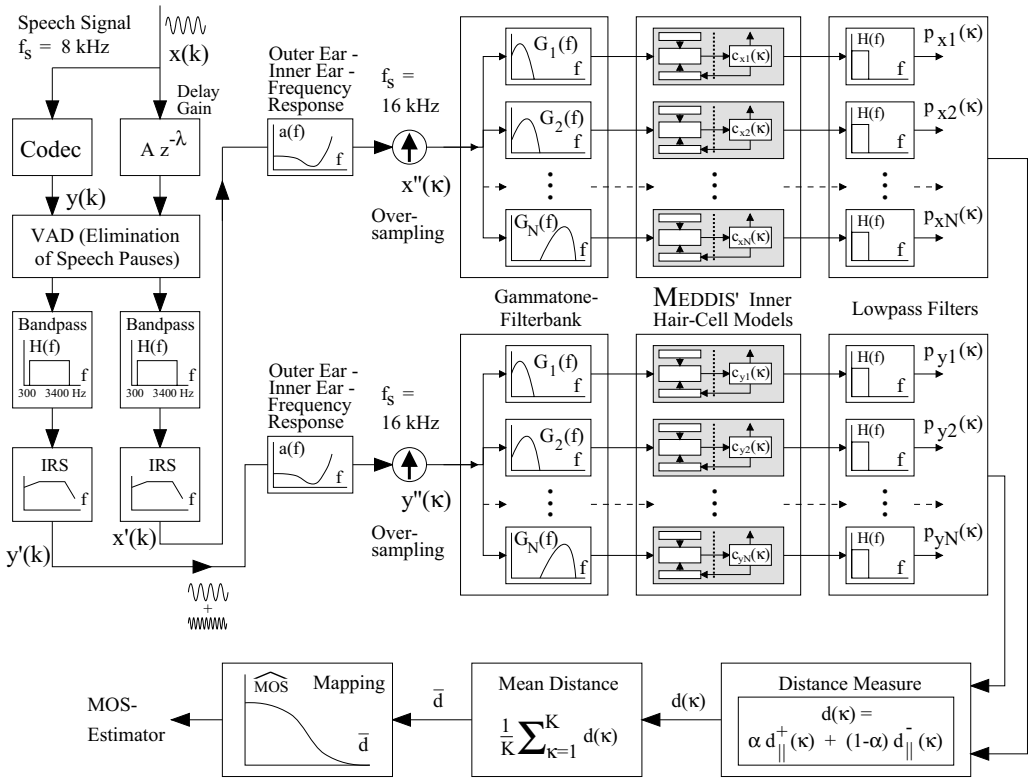


Figure 4: Objective speech-quality measure comparing nerve firing-patterns.

and subjective data thus not describing effects relevant to speech quality but solving a curve-fitting problem by parameter optimization.

The following distance measure $d(\kappa)$ proved to be successful:

$$d(\kappa) \;=\; \alpha\, d_{||}^{+}(\kappa) \;+\; (1-\alpha)\, d_{||}^{-}(\kappa)$$

with

$$d_{||}^{+}(\kappa) \;=\; \sum_{\nu} \max\{\,[p_{y\nu}(\kappa) - p_{x\nu}(\kappa)]\;,\,0\,\}$$

and

$$d_{||}^{-}(\kappa) \;=\; \sum_{\nu} \max\{\,[p_{x\nu}(\kappa) - p_{y\nu}(\kappa)]\;,\,0\,\} \qquad .$$

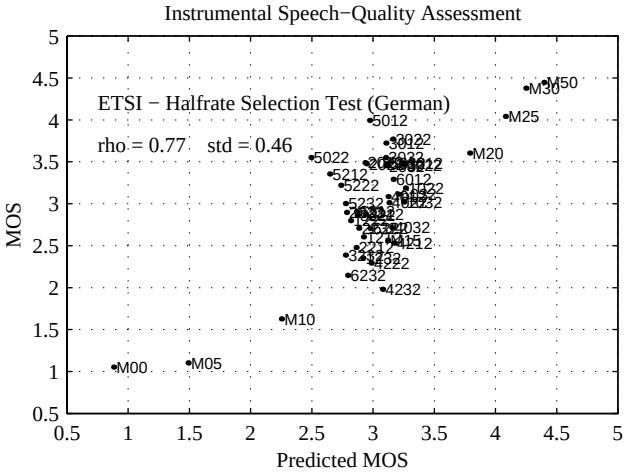
$d_{||}^{+}(\kappa)$ is the mean absolute error of the firing probabilities of $y''(\kappa)$ being higher than those of $x''(\kappa)$, and $d_{||}^{-}(\kappa)$ is the mean absolute error of the firing probabilities of $y''(\kappa)$ being lower than those of $x''(\kappa)$. $d(\kappa)$ is thus the weigthed sum of two distance measures $d_{||}^{+}(\kappa)$ and $d_{||}^{-}(\kappa)$, and the weighting is performed by the proper choice of the parameter α. The weighting of these two errors is strongly unsymmetric ($\alpha \sim 0.8$ ... 0.9) because signal components introduced by the codec are much more annoying than components which are attenuated or left out [5].

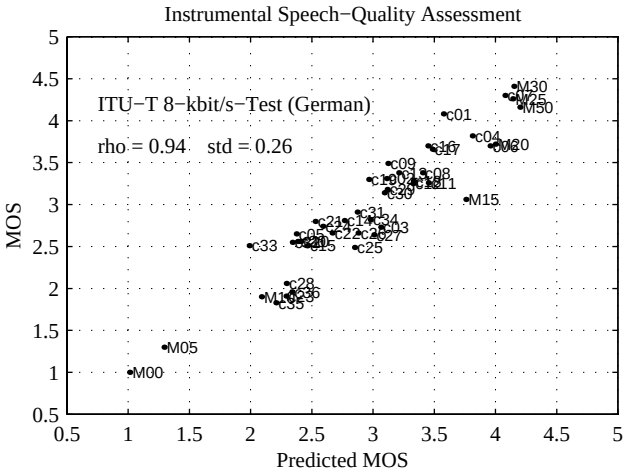
5. MAPPING

The values $d(\kappa)$ obtained for all time steps are averaged, and the mean distance $\bar{d}$ must be mapped to an estimator of the MOS since there should be a non-linear (but monotonic) relationship between $\bar{d}$ and the MOS (small distances should indicate good quality, i.e. high MOS). An arctangent function (the parameters are determined by an optimization procedure that aims to minimize the MSE between mean opinion scores and their estimations) provides an adequate mapping and can model threshold effects.

6. RESULTS

In Figures 5 and 6 four tests are given as examples. The subjective values (MOS or 'goodness') are plotted versus the objective scores obtained by an instrumental measure as described above. The correlations between the subjective and objective scores are quite high indeed for three tests. However, the predicted MOS-values for the ETSI Half-Rate Selection Test are less reliable (other measures using specific-loudness patterns have also difficulty in reproducing the subjective results of this test). All in all, the results are quite promising and encourage further research activities.





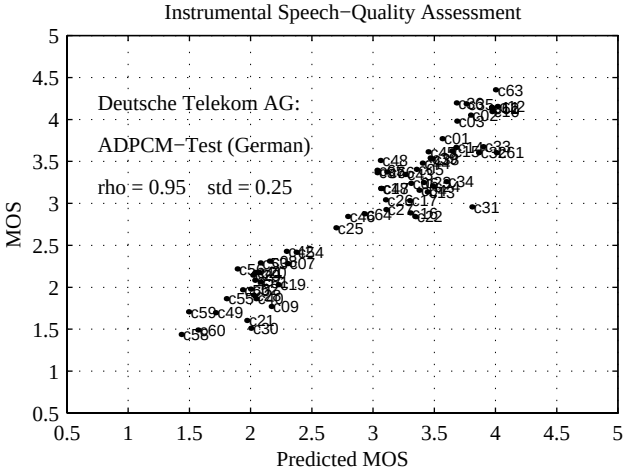


Figure 5: Application of the instrumental speech-quality measure to the ETSI Half-Rate Selection Test, the ITU-T Characterization Test of the 8 kbit/s-codec (G.729) and an ADPCM-Test conducted by Deutsche Telekom AG. The German speech material of these tests was used.

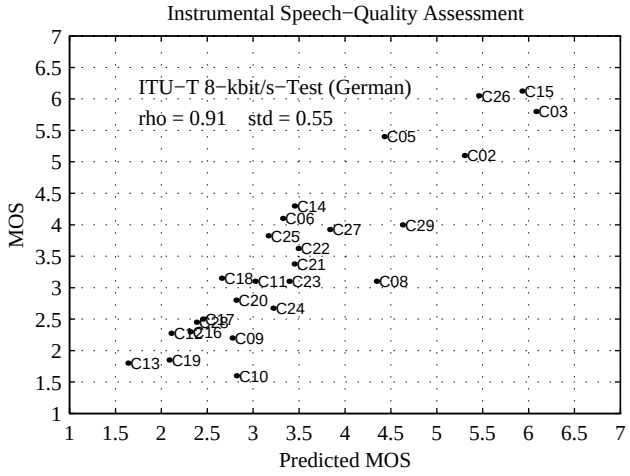


Figure 6: Application of the instrumental speech-quality measure to a speech codec test performed at Bochum University (German language). A 'goodness'-scale ranging from 0 to 10 was used instead of the MOS-scale. Unlike the other three tests, a slight modification of the distance measure (the relative weighted absolute distance was calculated) increased the correlation.

7. REFERENCES

[1] S. R. Quackenbush, T. P. Barnwell III und M. A. Clemens, Objective Measures of Speech Quality, Prentice Hall, 1988.

[2] Malcom Slaney, An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank, Apple Computer Technical Report #35, Apple Computer Inc., 1993.

[3] E. Zwicker and H. Fastl, Psychoacoustics - Facts and Models, Springer-Verlag Berlin Heidelberg, 1990.

[4] S. Wang, A. Sekey und A. Gersho, "Auditory Distortion Measure for Speech Coding", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 493-496, 1991.

[5] ITU-T, "Objective quality measurement of telephone-band (3400-3400 Hz) speech codecs", ITU-T Recommendation P.861, Geneva, 1997.

[6] John G. Beerends und Jan A. Stemerdink, "A Perceptual Speech-Quality Measure Based on a Psychoacoustic Sound Representation", J. Audio Eng. Soc., Vol. 42, No. 3, March 1994.

[7] Ray Meddis, "Simulation of mechanical to neural transduction in the auditory receptor", J. Acoust. Soc. Am., Vol. 79, pp. 702-711, March 1986.

[8] Ray Meddis, "Implementation details of a computation model of the inner hair-cell auditory-nerve synapse", J. Acoust. Soc. Am., Vol. 87, pp. 1813-1816, April 1990.

[9] M. J. Hewitt und R. Meddis, "An evaluation of eight computer models of mammalian inner hair-cell function", J. Acoust. Soc. Am., Vol. 90, pp. 904-917, August 1991.