# ON THE APPLICATION OF A PSYCHOACOUSTICALLY MOTIVATED SPEECH-QUALITY MEASURE IN CELP SPEECH-CODING

*Markus Hauenstein and Norbert Görtz*
Institute for Network and System Theory, University of Kiel, Germany
Tel: +49 431 77572 406, Fax: +49 431 77572 403
e-mail: ng@techfak.uni-kiel.de

## ABSTRACT

The crucial task in a CELP speech codec consists of finding the optimal excitation vector for the synthesis filter. This is usually done in an 'analysis-by-synthesis' structure by minimizing the mean-square error of the original and the coded/decoded speech frame. It is a common assumption that distance measures other than MSE and adapted to the human auditory perception should result in better speech quality. Such measures could be based on scientific results provided by psychoacoustics. However, due to the computational load there is no possibility to implement complex psychoacoustical models in real-time speech codecs and, for the time being, we are restricted to the MSE. Nevertheless, it is interesting to study the potential of psychoacoustic distance measures to improve speech codecs if complexity restrictions are neglected. This paper shows how a psychoacoustics-based distance measure can be integrated into a CELP codec, and the unexpected results are presented.

## 1 INSTRUMENTAL SPEECH-QUALITY ASSESSMENT

Speech-quality assessment deals with the determination of the best speech encoder/decoder (codec) in a group of candidate codecs. The time-consuming and expensive subjective way of finding out which codec sounds best consists of asking some people to listen to the speech codecs and to rate them. Usually a scale from 1 (poor quality) to 5 (good quality) is used. The scores concerning the different codecs are collected and a mean-opinion score (MOS) is calculated for each codec. Objective methods aim to replace these time-consuming and expensive subjective tests by an instrumental measure, i.e. a computer program.

Figure 1 shows the basic structure of an objective speech-quality measure which is comparing *loudness patterns* of the codec input signal $x$ and the output signal $y$. These representations of the speech signals in a 3-dimensional space (specific-loudness versus location of the excitation on the basilar membrane versus time) are more closely related to the human speech perception than the corresponding time signals or their linear spectral equivalents. Thus audible degradations should be more clearly expressed in the specific-loudness domain than in the time or frequency domain.

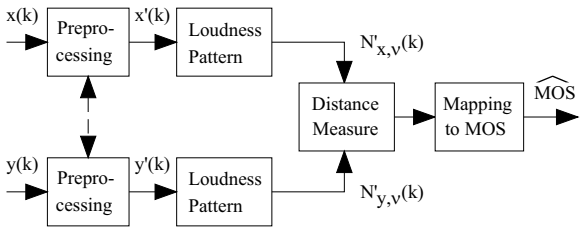


Figure 1: Basic structure of an instrumental speech-quality measure comparing loudness patterns

The basic idea of our specific measure was presented by Wang, Sekey and Gersho in [5]. However, our experiments with this type of measure only showed high correlations ($\rho \simeq 0.95$) between subjective and instrumental results after some major modifications concerning the signal preprocessing, the algorithm for calculating the loudness patterns and - very important - the distance measure for the comparison of the loudness patterns. Nevertheless, the way of calculating loudness patterns proposed in [5] is very efficient which was the reason for adopting it for our modified CELP codec.

## 2 LOUDNESS DENSITIES

Psychoacoustic experiments show that smearing effects take place in the cochlea while sound events are analyzed: Energy that is concentrated in the frequency domain (e.g. a pure tone) is exciting a whole section of auditory nerves on the basilar membrane in the inner ear. When using a sine or narrow-band noise, the location of the excitation maximum is a non-linear function of frequency or center frequency, respectively. This frequency smearing of energy can be modeled by a filter bank. The excitation of a single auditory receptor is assumed to be proportional to the output power of its corresponding

cochlea filter. Since the frequency responses overlap, a single sinusoid thus excites many receptors.

The frequency-to-location-transformation on the basilar membrane can be described by [3]

$$\frac{z}{\text{Bark}} = g(f)$$
$$= 13 \arctan\left(0.76\frac{f}{\text{kHz}}\right) + 3.5 \arctan\left[\left(\frac{1}{7.5}\frac{f}{\text{kHz}}\right)^2\right].$$

The variable $z$ has the unit 'Bark' and is almost linearly related to the perceived pitch of a sinusoid. The Bark scale can be regarded as a nonlinearly transformed frequency axis. On this $z$-axis, all cochlea filters $H_i(z) = H(z - z_i)$ have approximately the same shape and can be derived with sufficient accuracy by moving a prototype filter $H(z)$ to the location $z_i$ of the $i$-th auditory receptor.

Let $X_f(f,t)$ denote the short-time Fourier-spectrum of the input signal, and $Y_{f,i}(f,t)$ the spectrum of the output signal of the $i$-th filter $H_i(z)$ in the cochlea-filter bank. If we substitute $f$ by $z$, we have to demand

$$|X_f(f,t)|^2\, df \;=\; |X_z(z,t)|^2\, dz \qquad \text{and}$$

$$|Y_{f,i}(f,t)|^2\, df \;=\; |Y_{z,i}(z,t)|^2\, dz$$

for the power spectral densities (PSD). $X_z(z,t)$ and $Y_{z,i}(z,t)$ denote the corresponding power distributions on the Bark-scale.

Therefore, we find for the power $E_i(t)$ of $y_i(t)$, i.e. the excitation of the $i$-th auditory receptor at time $t$:

$$\begin{aligned} E_i(t) &= \int |Y_{z,i}(z,t)|^2\, dz = \int |H_i(z)|^2\, |X_z(z,t)|^2\, dz \\ &= \int |H(z - z_i)|^2\, |X_z(z,t)|^2\, dz \\ &= \int |H\,(g(f) - g(f_i))|^2\, |X_f(f,t)|^2\, df \\ &= \int W_i(f)\; |X_f(f,t)|^2\, df \end{aligned}$$

Thus we can estimate the excitation of a single auditory receptor at a given time $t$ by calculating an inner product of the input short-time PSD and the corresponding weighting function $W_i(f)$ of the receptor. As an example figure 2 shows an appropriate set of weighting functions for 20 auditory receptors. (In the implementation of the psychoacoustically modified CELP speech codec in the next section 128 receptors are used.) Before sound waves are analyzed by the nerve cells in the cochlea, they have to pass the outer and middle ear. We can model this transfer by a linear time-invariant filter. Each weighting function was pre-emphasized by the frequency response of this filter, therefore we find an 'envelope' in our set of weighting functions which corresponds to the frequency response from the outer ear to the inner ear.

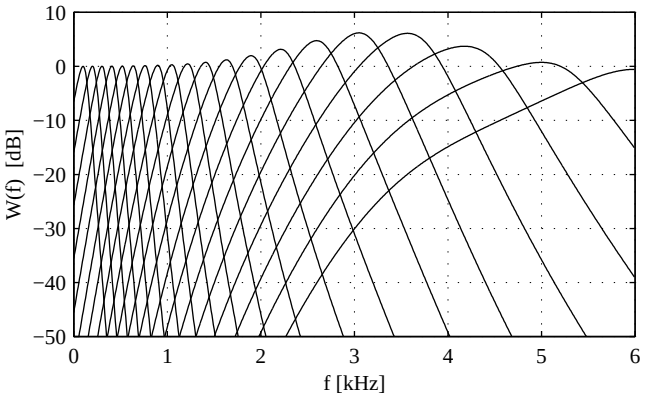


Figure 2: Estimation of the auditory-nerve excitation with PSD weighting functions for a set of 20 auditory receptors equidistantly distributed on the bark scale ($f_{min}$ = 0 Hz, $f_{max}$ = 6000 Hz)

The last step consists of applying a compression law derived by Zwicker [3] for the transformation from excitation $E_i(t)$ to specific loudness $N_i'(t)$:

$$N_i'(t) \;=\; 0.08 \left(\frac{E_{TQ,i}}{E_0}\right)^{0.23} \left[\left(\frac{1}{2} + \frac{1}{2}\frac{E_i(k)}{E_{TQ,i}}\right)^{0.23} - 1\right] \frac{\text{sone}}{\text{Bark}}$$

$E_0$ is the excitation corresponding to the intensity normalization value $I_0 = 10^{-12} W/m^2$, and $E_{TQ,i}$ describes the frequency-dependent excitation at the threshold of hearing.

## 3 MODIFIED CELP STRUCTURE

Most of the presently standardized speech codecs are of CELP-type [8]. Figure 4 shows the CELP analysis-by-synthesis structure which we used for our experiments. The codec processes frames of 160 narrow-band speech samples for which a 10-th order LPC-analysis [11] is carried out resulting in the coefficients for the synthesis filter. The frames are divided into four subframes with 40 samples each for which the search for the best excitation signal of the synthesis filter is performed. The excitation signal consists of the weighted sum of three shape-vectors: The first vector is called the 'adaptive' excitation and is constructed from previous filter inputs. The other two vectors are taken from two 'stochastic' codebooks containing noise sequences. The choice of the three vectors is successively optimized (according to the numbers at the dashed lines in figure 4) in a way that a given error criterion judging the difference between synthesized and original speech vector becomes minimal. The optimal gain-factors $v_a$, $v_{s1}$, and $v_{s2}$ minimizing the MSE for each shape-codevector from the codebooks are used to scale each of them while searching for the best candidate (as usual in CELP). The optimal gain-factors are also used for the reconstruction of the speech signal, i.e. $v_a$, $v_{s1}$, and $v_{s2}$ are not quantized in order to simplify the codec. The stochastic codebooks are untrained,

i.e. they consist of white noise sequences taken from a random generator. Surely, the performance of the codec could be improved by codebook-training [10]. Since the possible quality-gain by use of a better distance measure is to be investigated there is no need to spend much time on codebook training which would possibly require new algorithms because it depends on the distance measure.

In contrast to most present real-time implementations of CELP codecs we replaced the MSE-distance measure by a psychoacoustic measure with 128 receptors in the selection of the best shape codevectors. The scaling factors $v_a$, $v_{s1}$, and $v_{s2}$ are still calculated by analytically minimizing the MSE-criterion for each candidate shape codevector since the psychoacoustic distance measure can not be minimized analytically. The candidate speech signals are first transformed to the spectral domain via FFT and then converted to loudness densities by using the algorithm described in section 2. We select the codebook vectors which minimize the absolute error of the loudness densities between original and synthesized speech. Since the psychoacoustical distance measure requires the complete decoded signal (as received by the human ear) as input signal, some commonly used simplifications of the CELP-structure cannot be used. For instance the subtraction of the zero-input response of the synthesis filter (resulting from non-zero states from filtering the best excitation-signal of the previous frames) from the input speech to generate the target-vector for the codebook-searches is performed, but the zero-input response is "re-added" to the filtered and gain-scaled codevector to generate the tested candidate for the decoded speech-signal.

## 4 RESULTS

Figure 3 shows spectrograms giving an impression of the performance of our psychoacoustically modified CELP-codec. Unfortunately, we have to notice that the pitch structure is highly reduced in the synthesized speech which is also apparent when listening to the speech samples. We can explain this effect by the nature of the loudness densities: Since they result from smeared spectra, they do not carry much fine structure. So the correct synthesis of the spectral fine structure - i.e. pitch - is of no importance in the optimization procedure aiming to find the best excitation for the synthesis filter. This results in a "pitch-less" synthesized speech.

## 5 IMPROVEMENTS

We are thus not able to synthesize the perceptually very important spectral fine structure with an error criterion that compares loudness densities. In our CELP-codec, the synthesis of pitch is mainly the task of the adaptive excitation which is determined prior to the stochastic excitation vectors. A better pitch reconstruction can therefore be achieved by using different error criteria

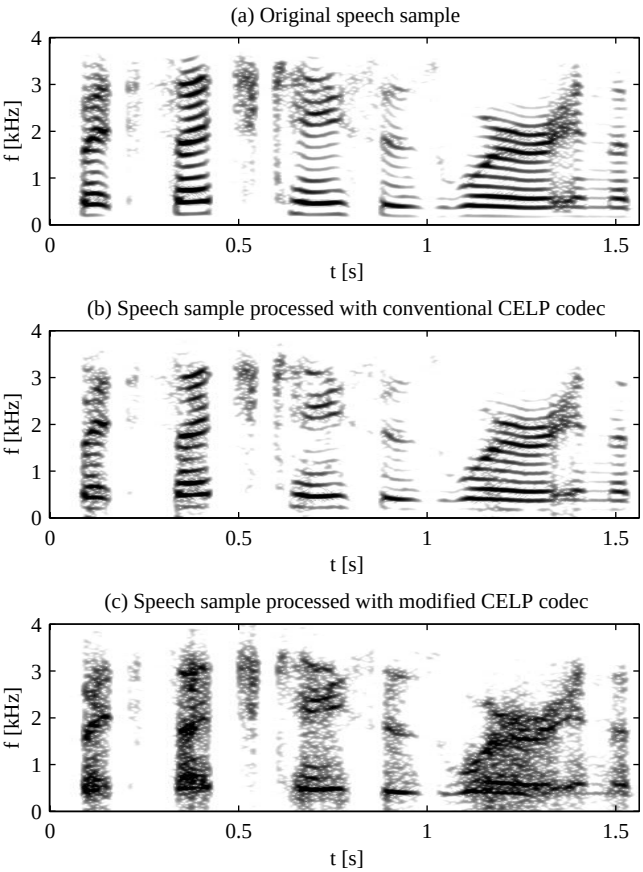


Figure 3: *Application of a psychoacoustically motivated distance measure to find an optimal excitation in a CELP speech codec: (a) Spectrogram of the input signal, (b) Spectrogram of the output signal if MSE is used as in conventional speech codecs (c) Spectrogram of the output signal of if a psychoacoustically motivated distance measure is used (calculation of auditory-nerve excitation (128 receptors) with DFT and weighting functions and nonlinear compression to specific loudness).*

for the selection of the adaptive and the stochastic excitation vectors. In a next step of our study we used the common MSE for the pitch regeneration and the loudness density difference only for the stochastic parts of the filter excitation. Although in informal listening tests the synthesized speech was judged to be equivalent in quality to speech synthesized by using only MSE, the SNR dropped by approximately 4 dB. Low bit-rate codecs like CELP typically reach SNR values of about 10 dB, and a degradation of this magnitude (4dB) is normally clearly audible if MSE is used as distance measure.

## 6 CONCLUSIONS

The substitution of the MSE by a psychoacoustically motivated error criterion as used in instrumental speech quality assessment did not result in better speech qual-

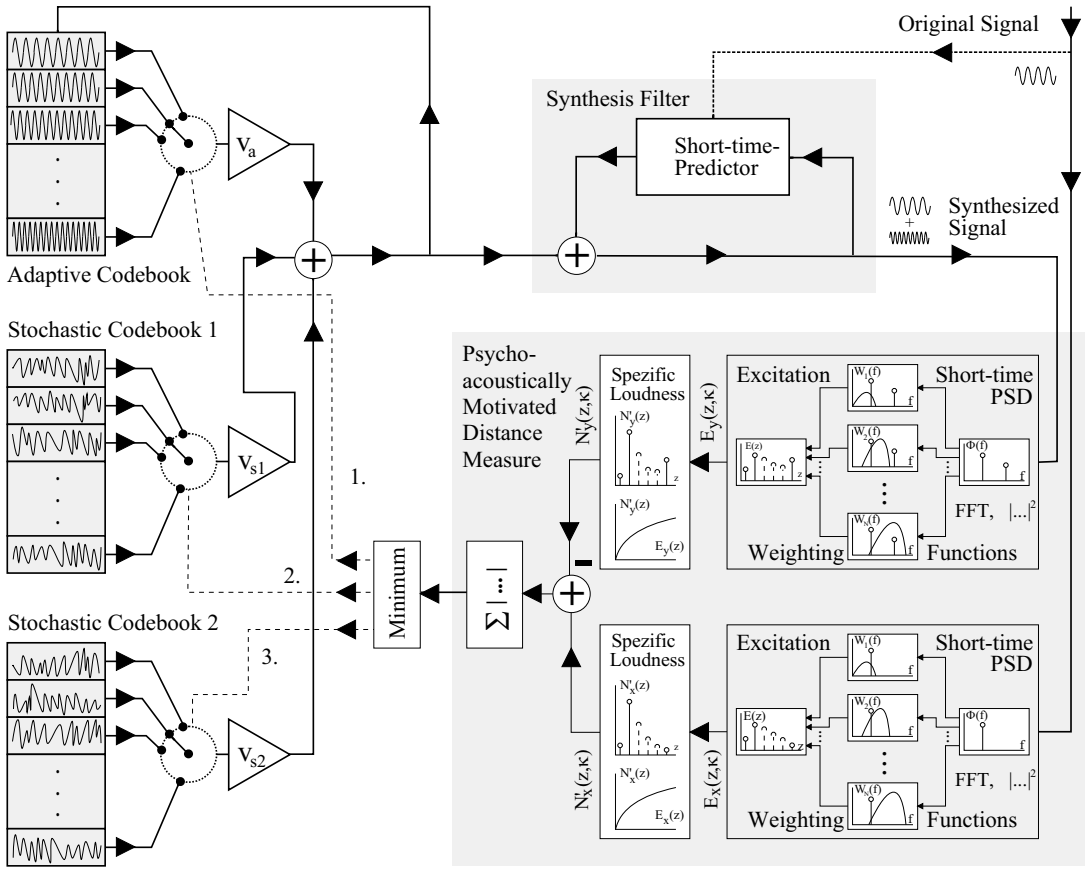


Figure 4: Psychoacoustically modified CELP codec

ity. On the contrary, the synthesized speech lacked pitch structure. This shows that speech-quality measures comparing loudness patterns are not sensitive enough to degradations in the spectral fine structure being introduced by a codec. Furthermore, the usage of the MSE in the adaptive part (and the new psychoacoustic distance measure in the stochastic part) of the excitation in the CELP codec achieved a speech quality equivalent to common CELP codecs using MSE in the adaptive and stochastic part as well. The SNR dropped greatly since the MSE (it's inverse is equivalent to SNR) was no longer minimized by parts of the CELP encoder. This shows that a CELP codec has the potential of synthesizing a set of perceptually equivalent speech samples, though we do not know if we have already reached the upper quality limit. This encourages further studies.

## 7 REFERENCES


[1] ITU-T Recommendation P.80, "Methods for subjective determination of transmission quality", *Telephone Transmission Quality*, Blue Book, Vol. V, Genf 1989.

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

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

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

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

[6] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *J. Acoust. Soc. Am.*, Vol. 87, pp. 1738-1752, April 1990.

[7] A. Sekey and B. Hanson, "Improved one-Bark bandwidth auditory filter", *J. Acoust. Soc. Am.*, Vol. 75, pp. 1902-1904, June 1984.

[8] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 937-940, 1985.

[9] M. Hauenstein, *Psychoakustisch motivierte Maße zur instrumentellen Sprachgütebeurteilung*, Dissertation, University of Kiel, Germany, 1997.

[10] Y. Linde, A. Buzo, R.M. Gray, "An Algorithm for Vector Quantizer Design", IEEE Trans. on Communications, Vol. COM-28, No. 1, pp. 84-95, January 1980

[11] J. Makhoul, "Linear Prediction: A Tutorial Review", Proceedings of the IEEE, Vol. 63, No. 4, April 1975, pp. 561-580