USE OF COMPUTATIONAL PSYCHOACOUSTICAL MODELS IN SPEECH PROCESSING: CODING AND OBJECTIVE PERFORMANCE EVALUATION

Jouni Koljonen   Matti Karjalainen

Helsinki University of Technology
Acoustics Lab., Otakaari 5 A
02150 Espoo 15, FINLAND

ABSTRACT

The primary objective of this study was to determine what benefit could be gained in speech coding by using a psychoacoustical frequency scale instead of a linear scale. To partially overcome the well known difficulties in objective speech quality measurements, a computational performance criterium based on psychoacoustical models was developed. Several Finnish phonemes were then coded using regular LPC and LPC computed on a psychoacoustically correct frequency scale (Bark scale)and the coding performance of these both methods was tested via computational performance tests. The results indicate a significant improvement in speech quality for the same bit-rate, when applying LPC on psychoacoustical frequency scale. Preliminary listening tests support both the better coding capability of the Bark LPC compared to the regular LPC and the reliability of the developed speech quality criterium as an objective performance evaluation method.

INTRODUCTION

Speech processing that is based rather on human perception than on conventional spectral representation has gained increasing interest during the last ten years, although the basic models representing human perception have been known for a much longer time. The use of psychoacoustical theories is justified due to the highly redundant nature of speech, from which the ear is able to filter out just the essential features. If speech is coded using conventional techniques (like LPC) the result cannot be psychoacoustically optimal due to this redundancy in the acoustical signal. To achieve a better result the coding algorithm should mimic ear in some way to remove the irrelevant data.

Speech quality measurement is another quite evident problem, where human perception models can be used. If a computational criterion yielding reliable measure of auditive difference between spectra could be developed, this would be a useful tool in performance evaluation of different coding techniques.

It was therefore the purpose of this study to investigate:

1. the implementation of models of human perception in speech coding based on speech production (LPC).
2. the development of a computational method based on human perception for objective performance evaluation of processed speech.

THE AUDITORY SPECTRUM: ITS BASIC THEORY AND COMPUTATION

In this paper the auditory spectrum will be regarded as an approximation of the equivalent of the linear power spectrum measured at the "output" of the peripheral auditory system. Those features that will be incorporated in this model do not deal with the mechanical to neural transduction properties of the ear at the basilar membrane level nor they do not include the processing that takes place at neural level. These features are omitted from this treatment not because they were insignificant, on the contrary, but because the mathematics involved would become too rigorous to be implemented practically.

The computation of the auditory spectrum will consist of three well known functional properties of the human ear:

1. Frequency weighting of the outer and middle ear.
2. Spectral warping from the linear (Hz) frequency scale to the psychoacoustic (Bark) scale.
3. The masking effect and the critical bandwidth phenomena.

Frequency weighting of the ear

This feature is easily included in the perception model by just a point by point multiplication of the original short-time Fourier transform of the time frame under consideration with a properly chosen equal loudness contour. We have used as a reference the 60 dB contour of the ISO recommandation R 226.

Spectral warping

Since the early work of Zwicker and Stephens, among others, various approximations of the dependence between the linear frequency scale and the psychoacoustical frequency have been proposed. We have chosen the Bark scale (1 Bark = 1 critical bandwith), which can be matemathically approximated with the equation (1) (1. Schröder):

$$f = 650*\sinh(x/7) \qquad (1)$$

where f is the frequency (kHz) and x is the critical band number (Bark).

Although equation (1) is rather simple, problems are encountered in digital calculations due to the unequal spaces between sample points of the warped spectrum, which prevents the use of the FFT algorithm. This can be avoided by performing an interpolation and a resampling of the warped spectrum.

CH1945-5/84/0000-0009 $1.00 © 1984 IEEE

Convolution with the basilar membrane spreading function

The process that leads to the critical band phenomena and masking effect in the ear can be regarded as a frequency domain convolution of the warped power spectrum and the basilar membrane spreading function (1. Schröder).

A block diagram which summarizes the computation of the auditory spectrum is shown in fig (1). The input to the process is the power spectrum on linear frequency scale of one frame of the analyzed time sequence. The output will be the auditory representation of this frame. It is worthwhile to notice that much information is lost during the processing yielding a smoothed version of the original spectrum. If, as we think, the model used is justified, this lost information must be considered as redundant.

The auditory spectrum as computed here is a more or less rude approximation of the real "auditive spectrum" due to the above mentioned shortcomings in representation of higher level processing and also due to the static nature of this model; neither forward nor backward masking has been considered. However it is still to be expected that the formulated auditory spectrum will be a useful signal processing tool in different problems, which deal with human perception of sounds.

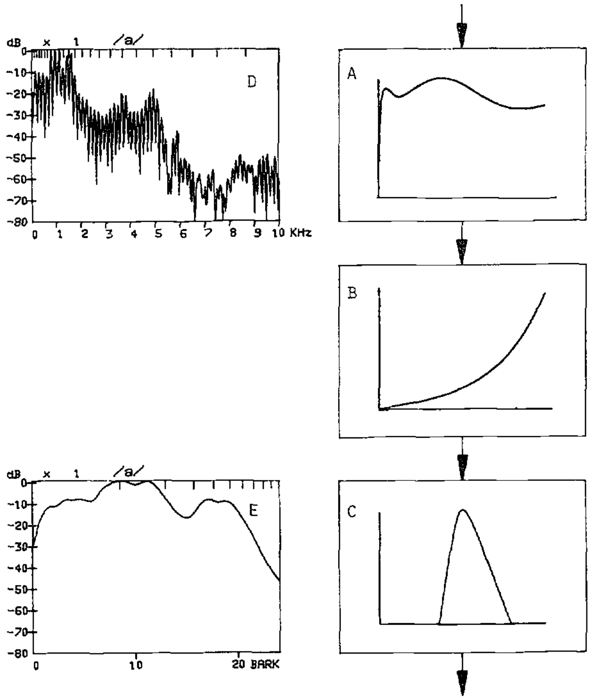

Fig. 1. A block diagram of the computation of the auditory spectrum. (A = equal loudness contour weighting, B = frequency warping(Hz > Bark), C = convolution with the basilar membrane spreading function, D = input to the auditory spectrum computation process (= linear power spectrum), E = output of the process (=auditory spectrum).

LPC COMPUTED ON THE BARK FREQUENCY SCALE

The Itakura-Saito error criterium, which is used in ordinary LPC, guarantees an uniform match across the linear frequency scale. It was previously pointed out that linear scale is not perceptually suitable and therefore ordinary LPC cannot give a psychoacoustically optimal result. This imperfectness can be avoided, if the linear spectrum is warped following the method described in the previous chapter and the LPC is applied to this warped spectrum. This representation will give an uniform match across the psychoacoustical frequency scale, thus giving an more optimal result.

The Bark scale has the property that higher frequencies are compressed relative to lower frequencies. The 0-5 Bark portion of the Bark scale is linear representing the 0-500 Hz portion of the linear scale. If LPC is applied on the linear and Bark scale in a way that the subbands 0-5 Bark and 0-500 Hz are coded with an equal number of bits the linear LPC will need more bits for every band larger than this, due to the logarithmic relationship between Bark and linear scale. Table 1 gives the bit-rate difference between Bark LPC and linear LPC for different bands.

Table 1.

Band	Percentage of bit-rate needed for Bark scale LPC
0-500 Hz	100 %
0-3000 Hz	46 %
0-6000 Hz	29 %
0-10000 Hz	21 %

For speech processing purposes the band of interest is about 0-5000 Hz and regarding this band the Bark LPC would theoretically need 35 % of the bit-rate needed for ordinary LP coding.

The above assumptions give a theoretical maximum advantage in a situation when the coding is performed with a Bark LPC degree just high enough to give a result that is perceptually not different from the original signal. In many practical situations it may be adequate that just the "phonetic" distance between coded and original speech is small enough to make the coded speech understandable. In this case the bit-rate values of table 1. must be increased. All the same, it is still to be expected that the decrease in bit-rate in favor of the Bark LPC method is of the order of 50 % (2. Makhoul).

The computation of the Bark LPC can be summarized in the following way:

1. Calculate the short-time spectrum of the time frame
2. Warp the linear spectrum to the Bark scale
3. Calculate the inverse Fourier transform yielding an autocorrelation like time domain signal.
4. Perform a LP coding on this signal.

It is clear that in the decoding procedure the above steps must be performed in reverse way and the warping fuction used in 2. must be inverted.

The Bark LPC method has also certain disadvantages, which do diminuish the above mentioned bit-rate advantage. Due to the highly nonlinear warping performed, the formants will also be distorted. Thus, the original ability in formant fitting that is partially behind the power of LP coding will be deteriorated. How much this disadvantage will reduce the overall properties of Bark LPC is rather difficult to evaluate.

Even a yet less redundant result could be achieved by coding the previously developed "auditory spectrum" instead of the warped spectrum. However this would lead to a huge amount of extra computation mainly due to the convolution operation. Also there would be some uncertainty about the function to be used in the deconvolution operation when decoding the signal. Perhaps the incorporation of just the non-linear frequency dependence will result in a adequate compromise between coding efficiency and mathematical manipulation.

In fig. 2 a typical example of the error signal between the auditory spectra of original and coded signals is plotted for both methods. The uniform fit of Bark LPC is quite evident as also are the large errors of ordinary LPC in the low frequency region.

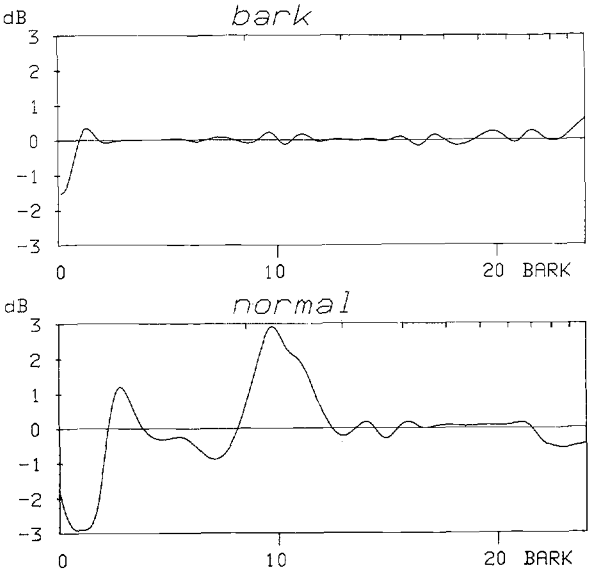

Fig. 2. The error in the auditory spectrum for Bark LP and ordinary LP coded test phoneme. (test phoneme /i/, LPC degree = 17, frequency band 0-10 kHz).

A CONFRONTATION BETWEEN BARK LPC AND ORDINARY LPC

In order to make clear the real coding peformance of LP coding applied on Bark frequency scale compared to ordinary LPC, several Finnish phonemes were coded with both methods and the previously developed representation based on the auditory spectrum was used as a test tool. Thus, the difference between the auditory spectrum of the coded and the original signal served as a criterium to judge the coding performance of the method in question. The difference between auditory spectra was calculated as a root-mean-square of the samples in the difference vector, and this value was properly multiplied in order to get the difference scaled on a bandwidth of 1 Bark.

The following Finnish phonemes pronounced by a male speaker were used: /a/,/ä/,/i/,/u/,/m/,/s/ and the frequency band was 0-10kHz. The band was chosen to be so wide in order to clearly find the benefit that can be achieved. The properly sampled and filtered test utterance was processed in order to get both the LPC and Bark-LPC spectra. The auditory spectra equivalents were then computed as also the auditory spectrum representation of the original signal. The above mentioned spectral difference measure between coded and original spectra yielded finally a score of the perceptual performance of both methods.

In fig. 3 a and b is shown the auditory spectral difference as a function of LPC degree for the Bark LPC (lower curve) and ordinary LPC (upper curve) for vowels and consonants respectively. (dBBS stands for dB (Bark)**0.5 due to the scaling performed.)

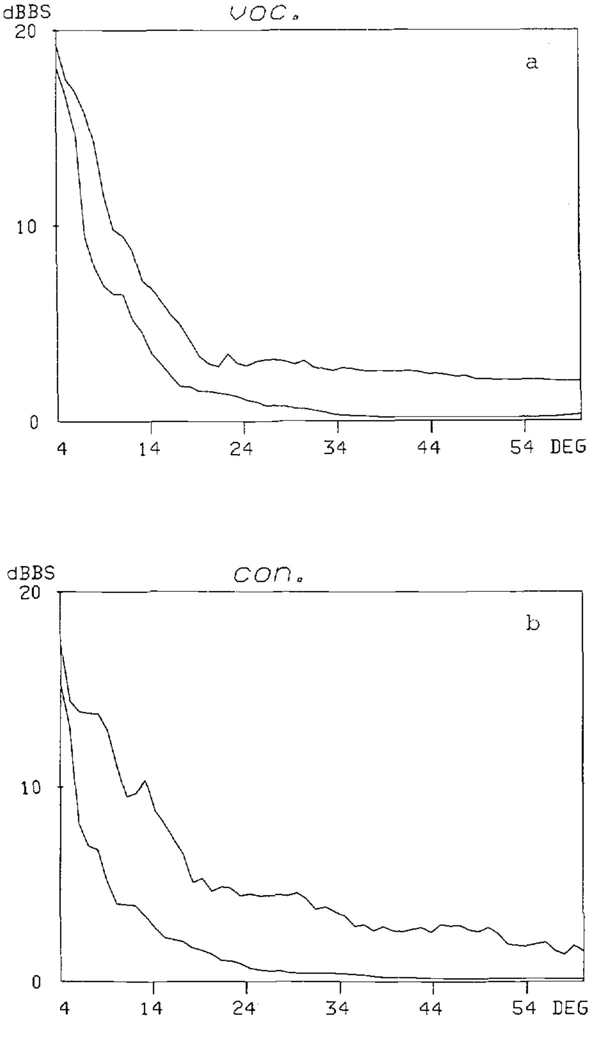

Fig. 3. The auditory spectral differences as a function of LPC degree for both coding methods (LPC on Bark frequency scale: lower curve, LPC on ordinary frequency scale: upper curve. a) Mean value curves for 4 different vowels: /a/,/ä/,/u/,/i/.), b) Mean value curves for 2 consonants : /s/,/m/)

The following deductions can be made:

1. Bark LPC yields a smaller auditory spectral difference for both vowels and consonants and across the whole abscissa
2. The superiority of the Bark LPC is more evident in coding of consonants. This is due to the fact that Bark LPC reserves more poles than ordinary LPC for the coding of low frequency zeros.
3. The typical "knee" in the curve for vowels (due to the formant fitting ability of LPC) is less pronounced for Bark LPC. Thus, there is some evidence that the power of Bark LPC is due more to overall fitting than formant fitting capability.
4. The convergence of ordinary LPC is slow which is probably caused by the difficulties in fitting the low frequency region.

In fig. 4 the ordinary LPC degree needed to achieve an auditory spectral difference equal to Bark LPC is plotted as a function of Bark LPC degree. Again vowels and consonants are considered separately. (The straight line is a result of regression analysis made on data collected from figs. 3 a and b.) It is normally assumed that a 22th degree ordinary LPC gives a quite satisfactory coding performance (on a 0-10kHz band). Thus the bit-rate reduction in favor of Bark LPC is in this case of the order of 40 % (vowels) and 60 % (consonants).

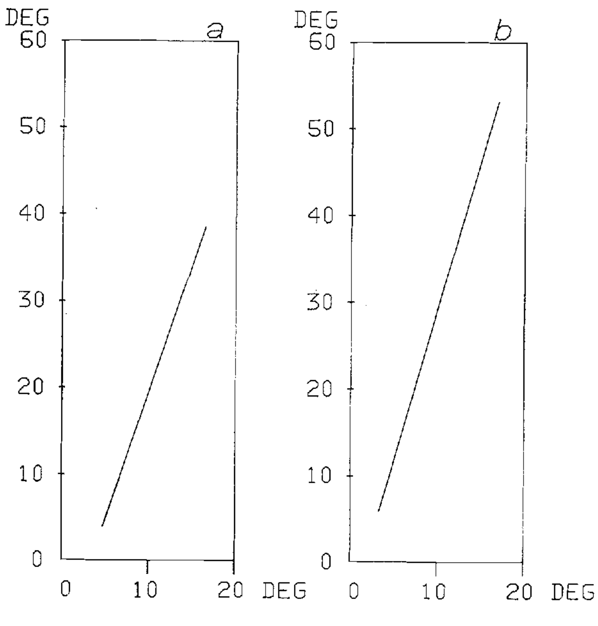

Fig. 4. LPC-analysis degrees that result in an equal coding performance measured with the auditory spectrum difference method for Bark scale and linear scale coding (vowels /a/,/ä/,/i/,/u/ and consonants /m/,/s/ are considered separately. a) vow., b) cons.).

Preliminary listening tests were also performed to test the reliability of the performance evaluation method based on the auditory spectrum and also to find out the real coding capability of Bark LPC. The coded and the original utterances were zero-phase-resynthesized from the data used in the above calculations. The listener had to judge if there were any differences between the coded and the original test phonemes. By changing LPC degree for following test signal pairs (coded-original) a threshold degree was determined which yielded a coding performance that perceptually equalled the original test signal.

In fig. 5 are shown the results, which quite well confirm the results based on the auditory spectrum. From figs. 3 a, b and 5 it can also be deduced that the threshold level is about 2.5 dBBS. This threshold level was found to be consistent for all phonemes, thus proving the reliability of the developed auditory measurement technique.

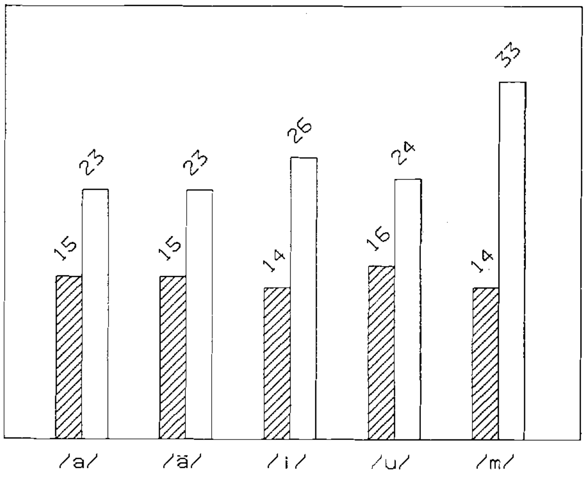

Fig. 5. The LPC degree needed to achieve a coded signal that is perceptually not different from the original test utterance. (Bark LPC: shadowed, ordinary LPC: white). Results based on preliminary listening tests.

CONCLUSIONS

A speech coding method based on LPC and psychoacoustically consistent prewarping of the spectrum was investigated. Also, a method for objective evaluation of coding performance was developed, and based on this method it is shown that the proposed coding technique is superior to ordinary LPC. Listening tests support both the coding capability of the proposed method as also the reliability of the objective performance evaluation method.

(1) Schröder, M. Atal, B. and Hall, L., Objective measure of certain speech signal degradations based on masking properties oh human auditory perception. In B. Lindblom and S. Öhman (Ed.) Frontiers of Speech CommunicationResearch. New York: Academic Press.

(2) Makhoul, J., and Cosell, L., LPCW : an LPC vocoder with linear spectral warping. Paper presented to the ICASSP-76.