AUDIO CODING USING A PSYCHOACOUSTIC PRE- AND POST-FILTER

Bernd Edler [1], *Gerald Schuller*

Multimedia Communications Research Laboratory
Bell Labs, Lucent Technologies, Murray Hill, NJ, U.S.A.
{bernd,schuller}@bell-labs.com,
http://www.multimedia.bell-labs.com

ABSTRACT

A novel concept for perceptual audio coding is presented which is based on the combination of a pre- and post-filter, controlled by a psychoacoustic model, with a transform coding scheme. This paradigm allows modeling of the temporal and spectral shape of the masked threshold with a resolution independent of the used transform. By using frequency warping techniques the maximum possible detail for a given filter order can be made frequency-dependent and thus better adapted to the human auditory system. The filter coefficients are represented efficiently by LSF parameters which can be adaptively interpolated over time.

First experiments with a system obtained by extending an existing transform codec showed that this approach can significantly improve the performance for speech signals, while the performance for other signals remained the same.

1. INTRODUCTION

Most of the current transform based audio coding schemes [1, 2] are designed in a way that a single spectral decomposition is used for both *irrelevancy* reduction and *redundancy* reduction. The irrelevancy reduction is obtained by a dynamic control of the quantizers for the individual spectral components according to perceptual criteria [3, 4]. This results in a temporally and spectrally shaped quantization error after the inverse transform in the corresponding decoder. The redundancy reduction is based on the decorrelating property of the transform. For audio signals with high temporal correlations this property leads to a concentration of the signal energy in a relatively low number of spectral components. By applying appropriate coding techniques, e.g. adaptive Huffman coding, this leads to a very efficient signal representation. The basic block diagram of a transform based audio encoder with a psychoacoustic model controlling the quantizers for the spectral components and the corresponding decoder is shown in Figure 1. This figure shows that in addition to the quantized spectral components, the quantizer control information needs to be transmitted.

One problem of this approach is the selection of the optimum transform length which is directly related to the frequency resolution. For relatively stationary signals a long transform with a high frequency resolution is desirable, since it allows to accurately shape the spectrum of the quantization error and furthermore provides a high redundancy reduction. For transients in the audio signal however a shorter transform has advantages due to its higher temporal resolution. This is necessary to avoid temporal spreading of quantization errors which may lead to an echoiness of the decoded signal.

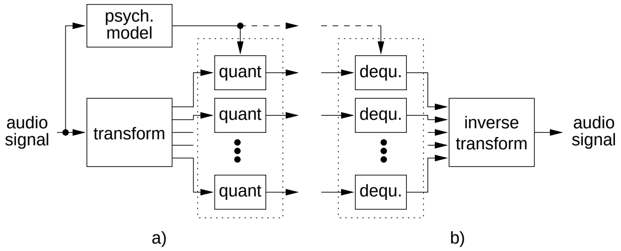

Figure 1: Basic block diagram of a transform audio encoder (a) and decoder (b).

Most solutions to this problem are based on the concept of dynamically switching the transform length dependent on the signal characteristics [5]. This approach is appropriate for dealing with the extreme cases where the signal is either relatively stationary or only contains transients. However, for signals containing different changes in separate frequency regions it is difficult to tradeoff between spectral resolution for the stationary components and temporal resolution for the transients. Speech signals are especially critical in this respect, particularly in vowels with changing formant structures, where the amplitudes of the higher harmonics can vary much faster than those of the lower harmonics.

Most CELP-type speech coders [6] overcome this problem by using a linear predictive scheme where the time varying spectral envelope of the input signal is approximated by the magnitude response of the LPC synthesis filter. The dynamic adaptation of the filter parameters allows a more flexible variation of spectral and temporal resolution than a transform. A spectral weighting filter is used in the encoder to select an excitation signal which leads to a spectrally shaped quantization error at the decoder output. Due to their specialization for speech signals and their limitation of the noise shaping functions to the encoder, CELP coders are generally not very well matched for coding of music signals. A similar approach is taken in so-called Transform Predictive Coding (TPC) [7, 8], where the residual signal

[1] Work was done while on leave from the University of Hannover, Germany

of an LPC analysis filter is transform coded.

Based on observations of the behavior of the different coding schemes, we propose a novel approach which separates the irrelevancy reduction, i.e. the spectral and temporal noise shaping, from the redundancy reduction. The goal is to combine the advantages of a psychoacoustic model, a noise shaping filter, and a transform in order to obtain a coding scheme suitable for speech as well as for music signals. Section 2 gives an overview of the concept how these components can be combined. The noise shaping pre- and post-filters are described in Section 3. Section 4 shows the mechanism to control the adaptation of the filter by a psychoacoustic model. Section 5 presents results obtained with an integration into an existing transform audio codec.

2. PSYCHOACOUSTIC PRE-FILTER

The new approach for perceptual audio coding presented here is to filter the signal before quantization and coding in the encoder. In the decoder a post-filter inverting the effect of the pre-filter is applied after decoding and de-quantization. If it is assumed that the distortions introduced by the quantization are additive white noise, the temporal and spectral structure of the noise at the decoder output is fully determined by the characteristics of the post-filter. The pre- and the post-filter are controlled by an appropriate psychoacoustic model. For this purpose the filter control information needs to be transmitted in addition to the quantized samples. The structure of an encoder with a pre-filter controlled by a psychoacoustic model and its corresponding decoder is shown in Figure 2.

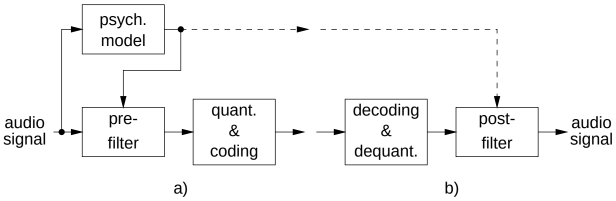

Figure 2: Encoder with psychoacoustic pre-filter (a) and corresponding decoder (b).

If a transform coding scheme is now applied to the pre-filtered signal, its spectral and temporal resolution can be fully optimized for achieving a maximum coding gain under pure MSE (mean square error) criteria, since the perceptual noise shaping is performed by the post-filter. This allows the use of fixed quantizer step-sizes, and no additional quantizer control information, e.g. individual scale factors for different regions of the spectrum, needs to be transmitted.

To adapt the filter characteristics to the masked thresholds (as generated by the psychoacoustic model) techniques known from speech coding can be used, where LPC filter parameters are used to model the spectral envelope of the speech signal.

The magnitude responses of the psychoacoustic post- and pre-filter should correspond to the masked threshold and its inverse respectively. Due to this similarity, known LPC analysis techniques can be applied with as major difference that now masked thresholds are used instead of short term spectra. Another difference is that for the pre- and post-filter not only the shape of the spectral envelope has to be taken into account, but the average level has to be included in the model by using appropriate gain factors.

One important advantage of the pre-filter concept over standard transform audio coding techniques is the greater flexibility in the temporal and spectral adaptation to the shape of the masked threshold.

It is of great advantage if the structure of the pre- and post-filter also supports the appropriate frequency dependent temporal and spectral resolution. Therefore a filter structure based on the so-called frequency-warping technique is used which allows filter design on a non-linear frequency scale.

3. STRUCTURE OF THE PRE- AND POST-FILTER

The most common forms of predictors use a minimum phase FIR filter in the encoder leading to an IIR filter in the decoder (Figure 3). This structure for the realization of a filter and its inverse has the advantage that it can be made time-varying quite easily, since the actual coefficients in both filters are equal and therefore can be modified synchronously.

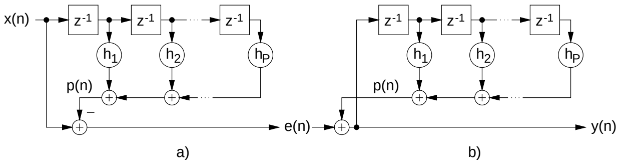

Figure 3: FIR predictor of order P (a) and its inverse (b).

For modeling masked thresholds a representation with the capability to give more detail at lower frequencies is desirable. For achieving such an unequal resolution over frequency the so-called frequency-warping technique [9] can be applied effectively. It is based on a principle which is known in filter design from techniques like lowpass-lowpass transform and lowpass-bandpass transform. In a discrete time system an equivalent transformation can be implemented by replacing every delay unit by an allpass. A frequency scale reflecting the non-linearity of the "critical band" scale [3] would be the most appropriate [10]. Generally the use of a first order allpass (Figure 4) already gives a sufficient approximation accuracy.

An implementable structure for this case is shown in Figure 5. Here delayless loops in the inverse filter are avoided by replacing the allpass sections by first order IIR sections [9]. Its coefficients g_k $(0 \leq k \leq P)$ are obtained from the original coefficients with the transformation

$$g_k = \sum_{n=k}^{P} C_{kn} h_n \text{ with } C_{kn} = \binom{n}{k} (1-a^2)^k (-a)^{n-k} \ . \quad (1)$$

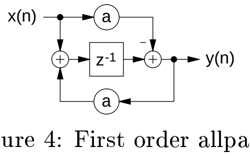

Figure 4: First order allpass.

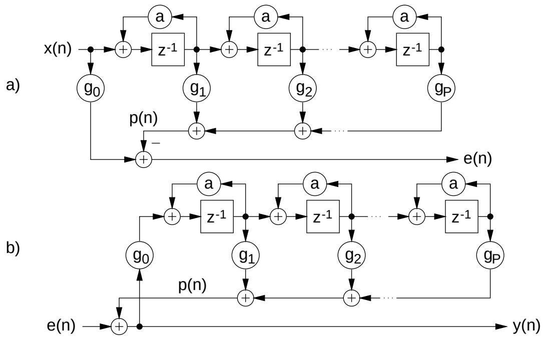

Figure 5: Structure of a filter (a) and its inverse (b) with frequency warping.

The use of a first order allpass leads to the following mapping of the frequency scale:

$$\tilde{\omega} = \omega + 2\arctan\frac{a\sin\omega}{1 - a\cos\omega}\ . \qquad (2)$$

The derivative of this function

$$\nu(\omega) = \frac{\partial\tilde{\omega}}{\partial\omega} = \frac{1 - a^2}{1 + a^2 - 2a\cos\omega} \qquad (3)$$

indicates whether the frequency response of the resulting filter appears compressed ($\nu > 1$) or stretched ($\nu < 1$). The warping coefficient a should be selected depending on the sampling frequency, e.g. at 32 kHz a value around 0.5 is a good choice for the pre-filter application. The advantages of the warping technique in the approximation of masked thresholds are shown in Section 5.

4. ADAPTATION AND CODING OF FILTER PARAMETERS CONTROLLED BY A PSYCHOACOUSTIC MODEL

Most psychoacoustic models generate thresholds with a fixed sample interval in time. Therefore the modeling approach can be separated into the approximation of the spectral shape of a single threshold and the tracking over time. In both steps, which are addressed in the following, the availability of efficient coding techniques is very important in order to limit the amount of side information.

4.1. Approximation of the Spectral Shape

In LPC analysis the auto-correlation function (ACF) is used in the calculation of optimum predictor coefficients. For the calculation of coefficients for the psychoacoustic pre-filter a pseudo-ACF is derived by applying an inverse DFT to samples of the masked threshold. For a regular FIR filter uniform sampling along the frequency axis is needed, whereas for a warped filter a sampling according to the non-linearity has to be applied.

Since the filter coefficients need to be transmitted to the decoder in order to control the post-filter, an efficient representation is very important. For this purpose a technique used in many modern linear prediction based speech coders proved to be very successful. It consists of a vector quantization (VQ) for line spectral frequencies (LSF) parameters. These techniques can be directly applied to the pre- and post-filter by converting warped LPC coefficients to (warped) LSF parameters and generating an appropriate codebook. Due to the fact that masked thresholds generally have a smoother behavior than short term spectra and due to the frequency warping the efficiency is even higher than in the original applications. This means that less bits are required for representing masked thresholds with a desired accuracy than for representing short term spectra.

As mentioned above another difference in the pre- and post-filter scheme compared to regular LPC is the additional transmission of an overall gain which represents the average of the masked threshold over frequency.

4.2. Approximation of the Temporal Shape

The temporal resolution required for modeling a time-varying masked threshold is mainly determined by the duration of the interval in which the human ear has a reduced ability to detect quantization noise preceding an onset of a masking tone. Since this so-called pre-masking spans only a few milliseconds, the model of the masked threshold should provide a resolution in the same range, i.e. $2\ldots4$ ms. Therefore without other mechanisms the filter parameters would need to be transmitted very frequently. However audio signals tend to be relatively stationary and such a high update rate would be a waste of transmission bandwidth.

Speech coders usually overcome this problem by transmitting filter parameters at a significantly lower rate (e.g. every $20\ldots30$ ms) and interpolating in-between. This is appropriate for this type of codecs since they only model the shapes of the spectral envelopes but not the signal energy, which is usually transmitted more frequently, e.g. as excitation gain information. Fast changes of masked thresholds, as they can occur at attacks of percussive instruments, however could not be modeled accurately enough with such a fixed interpolation scheme. Therefore a more flexible interpolation scheme is used, in which the transmission frequency of the filter parameters is adapted to the variation of the masked threshold. The transmission time instances need to be signaled by additional control data.

An overview on the procedure for pre- and post-filter adaptation is given in Figure 6.

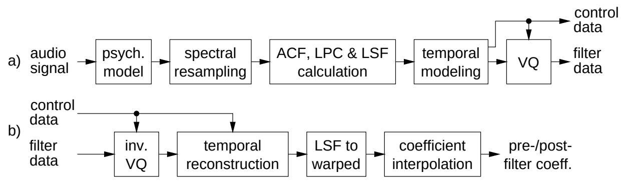

Figure 6: Adaptation of pre- and post-filter, a) modules only needed in the encoder, b) modules needed in encoder and decoder.

5. RESULTS

The psycho-acoustic pre- and post-filter were integrated in the PAC audio coder [1], such that only its transform and lossless coding parts were used.

First experimental evaluations focused on the masked threshold modeling capabilities of the pre- and post-filter technique. For this purpose, frequency responses of the post-filter resulting from the LPC model were compared to the masked thresholds obtained from the psychoacoustic model. Figure 7 shows for one analysis frame of a male speech signal the short term spectrum, the masked threshold from the psychoacoustic model, and the frequency response of a 12-th order LPC-like post-filter. It shows a very high approximation accuracy in the upper half of the frequency band due to the smooth behavior of the masked threshold. Towards the lower band edge however much of the detail in the masked threshold gets lost.

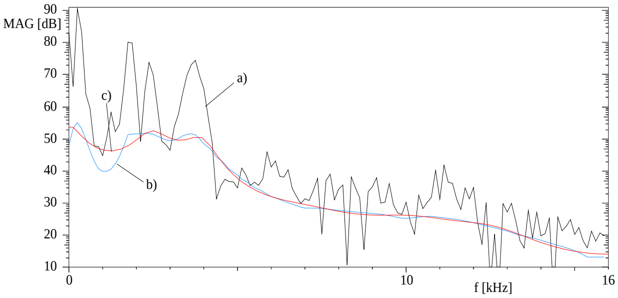

Figure 7: Short term spectrum of a male speech signal sampled at 32 kHz (a), masked threshold from psychoacoustic model (b), and frequency response of a 12 order LPC-like post-filter (c).

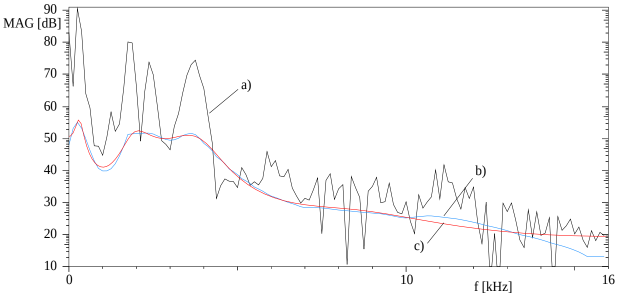

Figure 8: Short term spectrum of a male speech signal, masked threshold from psychoacoustic model (b), and frequency response of a 12 order post-filter with frequency warping coefficient $a = 0.5$ (c).

The interesting question is now, how much this uneven behavior can be reduced by using filters with frequency warping. Figure 8 shows the same spectrum and masked threshold as Figure 7. The third curve however was obtained from a warped filter structure according to Figure 5 with a warping coefficient $a = 0.5$. This frequency response provides clearly a better approximation at low frequencies. It comes at the cost of a slightly higher deviation at the upper band edge. The subjective effect of this deviation however can be assumed to be relatively small when the critical band width is taken into account.

Informal subjective evaluations comparing both structures integrated in a complete coding system based on PAC indicated a clearly audible improvement with the warped filter. This coding system was also used for further informal subjective evaluations with bit rates ranging from 16 to 24 kbit/s. Using informal listening tests, the new system was compared to PAC in its original form. The general observation was that for speech signals the pre-filter technique led to a clear improvement in quality, while for music signals there was no significant performance difference.

6. CONCLUSION

A novel concept for perceptual audio coding is presented based on the addition of a pre- and post-filter controlled by a psychoacoustic model to a transform coding scheme.

First experiments with a system obtained by extending an existing transform codec already showed promising results. Especially for speech signals, which are amongst the most critical signals, clear improvements are obtained.

Another interesting aspect of the pre-filter concept is that it is not restricted to the coding scheme PAC provides, but it can also be used in combination with other (lossless) coding methods, to obtain perceptually lossless coders.

7. REFERENCES

[1] D. Sinha, J. D. Johnston, S. Dorward, and S. Quackenbush, “The perceptual audio coder (PAC),” in *The Digital Signal Processing Handbook* (V. Madisetti and D. B. Williams, eds.), ch. 42, Boca Raton, Florida: CRC Press, IEEE Press, 1997.

[2] K. Brandenburg and M. Bosi, “Overview of MPEG audio: Current and future standards for low bit rate audio coding,” *J. Audio Eng. Soc.*, vol. 45, pp. 4–21, Jan./Feb. 1997.

[3] M. R. Schroeder, B. S. Atal, and J. L. Hall, “Optimizing digital speech coders by exploiting masking properties of the human ear,” *Journal of the Acoust. Soc. Am.*, vol. 66, pp. 1647–1652, Dec 1979.

[4] J. H. Hall, “Auditory psychophysics for coding applications,” in *The Digital Signal Processing Handbook* (V. Madisetti and D. B. Williams, eds.), pp. 39–1:39–22, CRC Press, IEEE Press, 1998.

[5] B. Edler, “Coding of audio signals with lapped transforms and adaptive window functions (in German),” *Frequenz*, vol. Band 43, Nr. 9, pp. 252–256, 1989.

[6] W. B. Kleijn and K. K. Paliwal, “An introduction to speech coding,” in *Speech Coding and Synthesis*, Amsterdam: Elsevier, 1995.

[7] J.-H. Chen and D. Wang, “Transform predictive coding of wideband speech signals,” in *IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 275–278, 1996.

[8] S. A. Ramprashad, “A multimode transform predictive coder (MTPC) for speech and audio,” in *IEEE Speech Coding Workshop*, pp. 10–12, June 1999.

[9] H. W. Strube, “Linear prediction on a warped frequency scale,” *J. of the Acoust. Soc. Am.*, vol. 68, pp. 1071–1076, 1980.

[10] U. K. Laine, M. Karjalainen, and T. Altosaar, “Warped linear prediction (WLP) in speech and audio processing,” in *IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. III–349 – III–352, 1994.