Perceptual Audio Coding Using a Time-Varying Linear Pre- and Post-Filter

Bernd Edler[1], Christof Faller, Gerald Schuller

Multimedia Communications Research Laboratory
Bell Labs, Lucent Technologies
700 Mountain Ave., Murray Hill, NJ 07974, USA
{bernd|cfaller|schuller}@bell-labs.com
[1]was on leave from the University of Hannover, Germany

Recently, a new concept for perceptual audio coding was presented, which is based on a pre-filter in the encoder and a corresponding post-filter in the decoder, both controlled by a psychoacoustic model. It enables individual selection of spectral and temporal resolutions for irrelevancy reduction and redundancy reduction. This paper addresses problems related to the efficient transmission of the filter parameters and presents techniques for efficient temporal and spectral modeling of masked thresholds using linear time-varying filters.

1 INTRODUCTION

Most transform based audio coders operating at bit rates around or below 1 bit/sample produce audible artifacts when applied to quasi-stationary signals such as speech. The decoded speech signal often has reverberant sounding artifacts. This problem can be largely explained by the fact that these coders use the transform for both irrelevancy removal and redundancy reduction, and that usually only 2 different transform sizes are used.

The irrelevancy reduction exploits masking properties of the human auditory system. It can be performed by applying individual quantizers to the spectral components, controlled by a psychoacoustic model. At the output of the inverse transform in the decoder this results in a quantization error temporally and spectrally shaped according to perceptual criteria [1, 2].

The redundancy reduction makes use of the energy compaction property of the transform, which concentrates the energy of signals with high temporal correlations in a relatively low number of spectral components. This effect can then be exploited by adaptive bit allocation and appropriate coding techniques.

While such a coding technique has the advantage that its basic structure is relatively simple (Figure 1) [3, 4], the use of a single spectral decomposition restricts the flexibility for an individual optimization of irrelevancy and redundancy reduction. The biggest problem is the selection of the optimum transform length. For stationary signals a big transform length corresponding to a high frequency resolution is required, since it enables accurate spectral noise shaping and performs a high energy compaction. For non-stationary signals, however a transform length is required which is small enough to provide sufficient temporal resolution for adapting the shape of the quantization noise.

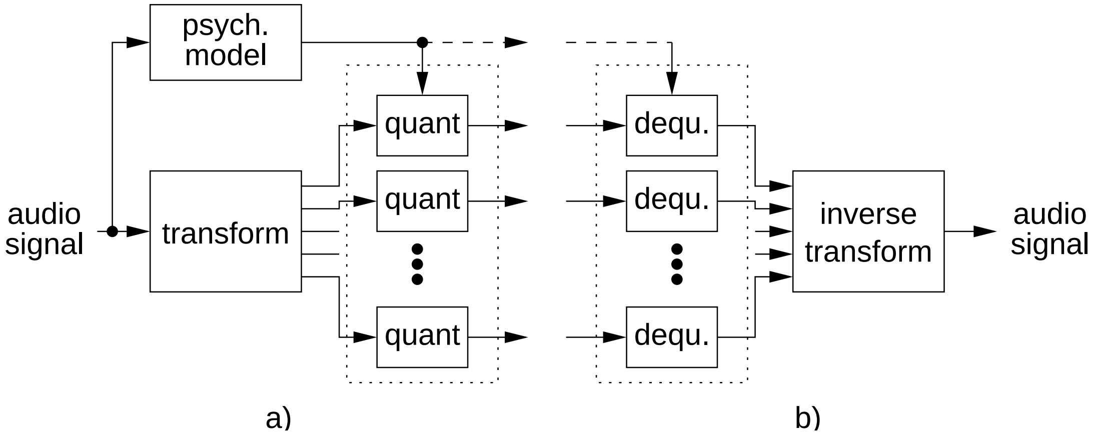

Fig. 1. Simplified block diagram of a transform audio encoder (a) and decoder (b).

The new coding technique presented here is based on the idea of using separate units for irrelevancy and redundancy reduction [5]. In the following it is shown how this can be achieved by using adaptive filters for the quantization noise shaping in combination with transform coding techniques optimized for redundancy reduction. The basic principle is introduced in Section 2, the use of “frequency warping” techniques is described in Section 3. Special attention is paid to the adaptation mechanisms in Section 4 and to the coding of filter parameters in Section 5. Results of the implementation in a complete audio coder are presented in Section 6.

2 TIME-VARYING PRE- AND POST-FILTER FOR AUDIO CODING

The main task of irrelevancy reduction in an audio coder is quantization noise shaping according to perceptual criteria.

Let us first assume the ideal case of a *transparent* coder, which would achieve the minimal bit rate without audible differences compared to the original signal. In this case the quantization noise would have a spectral and temporal shape matched to the masked threshold generated by the audio signal itself.

The usual rate-distortion theory on the other hand treats the case of minimizing the mean squared error or maximizing the SNR of the reconstructed signal, for a given bitrate. This leads to a quantization noise which is very similar to white noise, i.e. it has a nearly flat power spectral density.

Now we want to design a perceptual coder in which the transform is only used for the redundancy reduction part (or in a mean squared error sense). The quantization noise shaping is implemented by filtering the output signal of the inverse transform in the decoder. This filtering is done such that the quantization error becomes shaped according to the masked threshold. In order to maintain the shape of the audio signal, a corresponding inverse filter has to be applied prior to the transform in the encoder. This pre-filter in the encoder therefore must be controlled by a psychoacoustic model in a way that it approximates an inverse to the time-varying masked threshold. The block diagrams of an audio encoder using such a pre-filter and a decoder with the corresponding post-filter are shown in Figure 2.

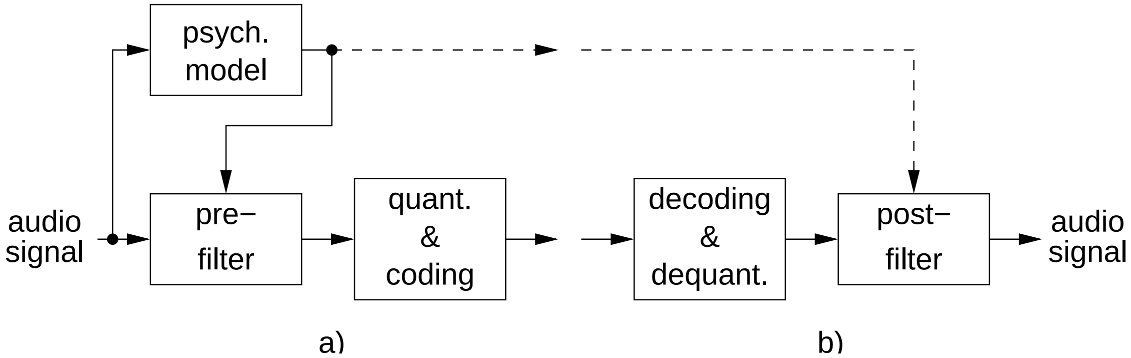

Fig. 2. Encoder with psychoacoustic pre-filter (a) and decoder with corresponding post-filter (b).

The frequency resolution of the transform can now be selected for achieving a maximum redundancy reduction and does not need to take into account perceptual effects like temporal masking. Thus the stages for irrelevancy reduction and redundancy reduction become separated and each stage can be optimized independently. This has clear advantages for signals which are neither stationary nor non-stationary, such as speech.

For the control of the pre- and post-filter by a psychoacoustic model techniques known from linear predictive coding (LPC) can be adapted. Predictors are used for exploiting correlations in the input signals, i.e. for redundancy reduction.

2.1 Linear Prediction

The following is a short description of the well know linear prediction technique. We will use this technique to construct our pre- and post-filter.

The predictor in the encoder produces from previous samples a predicted value which is as close to the input sample as possible. If $p(n)$ is the predicted value, and $x(n)$ is the input signal,

$$p(n) = \sum_{p=1}^{P} a_p x(n-p) \ . \tag{1}$$

A coding gain then is achieved by transmitting the difference, i.e. the prediction error signal

$$e(n) = x(n) - p(n) \ , \tag{2}$$

which can be coded more efficiently than the input signal. The predictor in the decoder has to re-generate the predicted value from the transmitted prediction error so that without quantization the sum of both would result in a reconstruction of the input signal (Figure 3). In the prediction error signal ideally no correlations would be left, which corresponds to white noise. The prediction and summation in the decoder reconstructs

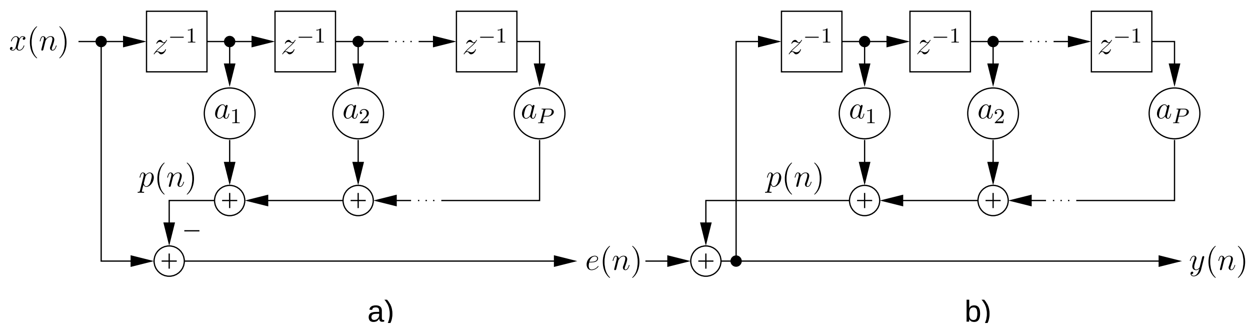

Fig. 3. Linear FIR predictor of order P (a) and its inverse (b).

the signal with its original spectral shape. This means that the inverse predictor forms a filter with a frequency response which approximates the signal spectrum.

A predictor with a finite impulse response (FIR) however only can approximate this behavior. The coefficients a_p for the optimum FIR predictor of a given order P can be derived from the auto-correlation function (ACF) $s_{xx}(n)$ of the input signal by solving the linear equation system

$$\sum_{p=0}^{P-1} s_{xx}(|p-n|)a_p = s_{xx}(n+1)\ , \quad 0 \le n < P\ . \tag{3}$$

The calculation of the predictor coefficients therefore consists of the two steps for estimating the ACF and solving equation 3 as shown in Figure 4a.

2.2 Psycho-Acoustic Pre- and Post-Filter

For the pre- and post-filter we now want to use the filter structure of the predictor, but approximate masked thresholds instead of signal spectra. If the psychoacoustic model describes the masked threshold as a power spectral density function ($M(\omega)$), this can be converted into a function corresponding to an ACF ($s_{mm}(n)$) of a signal with the same power spectral density using an inverse Fourier transform. In practical applications the masked threshold is sampled at equidistant frequencies and converted with an inverse Discrete Fourier Transform (DFT). Compared to the calculation of predictor coefficients the pre-filter parameters are obtained with the procedure shown in Figure 4b.

Another difference to LPC is that the pre-and post-filter not only need to model the spectral shape but also the total power of masked noise. Therefore the filter structure of Figure 3 has to be extended by a gain factor which is controlled by the integral of $M(\omega)$ which is equal to $s_{mm}(0)$.

3 PRE- AND POST-FILTER WITH FREQUENCY WARPING

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.

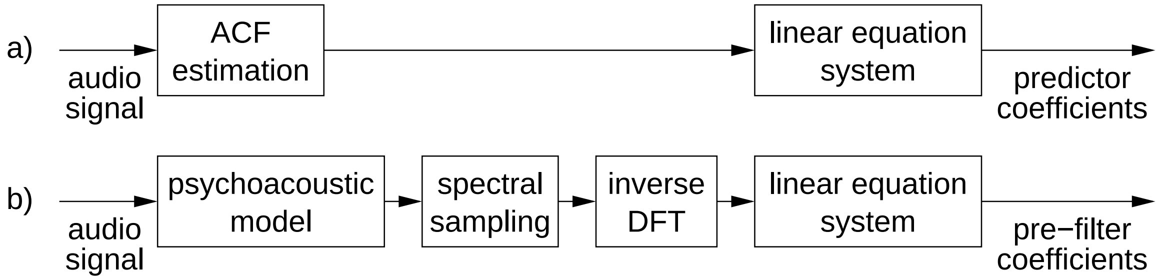

Fig. 4. Comparison of the steps for calculation of predictor coefficients in traditional LPC (a) and filter coefficients for our psycho-acoustically controlled pre- and post-filter (b).

A very important characteristic in the temporal behavior of masked thresholds is a relatively short rise time starting even before the onset of a signal (masker) and a longer decay after it is switched off.

Some important spectral properties of masked thresholds on the other hand can be observed in the presence of stationary single tone maskers. The masked threshold is spread around the masker frequency, sloping off slower towards higher frequencies than towards lower frequencies. On a linear frequency scale the steepness of both of the slopes strongly depends on the masker frequency, leading to flatter slopes with increasing masker frequency. However, on the non-linear so-called Bark scale, the shapes of masked thresholds are almost frequency independent [1].

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 the so-called frequency-warping technique [6] is applied. It allows filter design on a non-linear frequency scale based on a principle known 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. This procedure maps the frequency response of a filter to a non-linear frequency scale. If the allpass system is designed properly, the amount of spectral details which can be approximated by the filter structure can have a frequency dependency very similar to that of the frequency dependent resolutions of masked thresholds.

A first order allpass as shown in Figure 5 already provides a sufficient amount of frequency warping, as several studies in the field of frequency warped prediction have shown [6] [7].

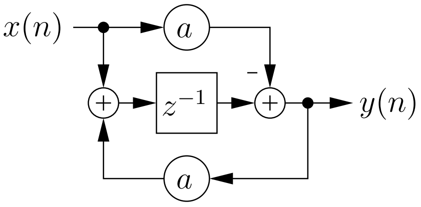

Fig. 5. First order allpass.

However, the direct substitution of the allpass (Figure 5) into the FIR predictor structure is only possible for the pre-filter (Figure 3a) and not for the post-filter. Since the allpass has a direct path without delay from its input to the output, its substitution into the feedback structure according to Figure 3b would result in a zero-lag loop. Therefore a modification of the filter structure is required.

In order to overcome the zero-lag problem, the delay units of the original structure can be replaced by first order IIR filters containing only the feedback part of the allpass structure [6]. However the filter coefficients need to be modified to obtain the same frequency response as a structure with allpass units. The coefficients g_p of the new structure (Figure 6) can be obtained from the original warped coefficients a_p with a transformation described in [6], which also introduces an additional coefficient g_0.

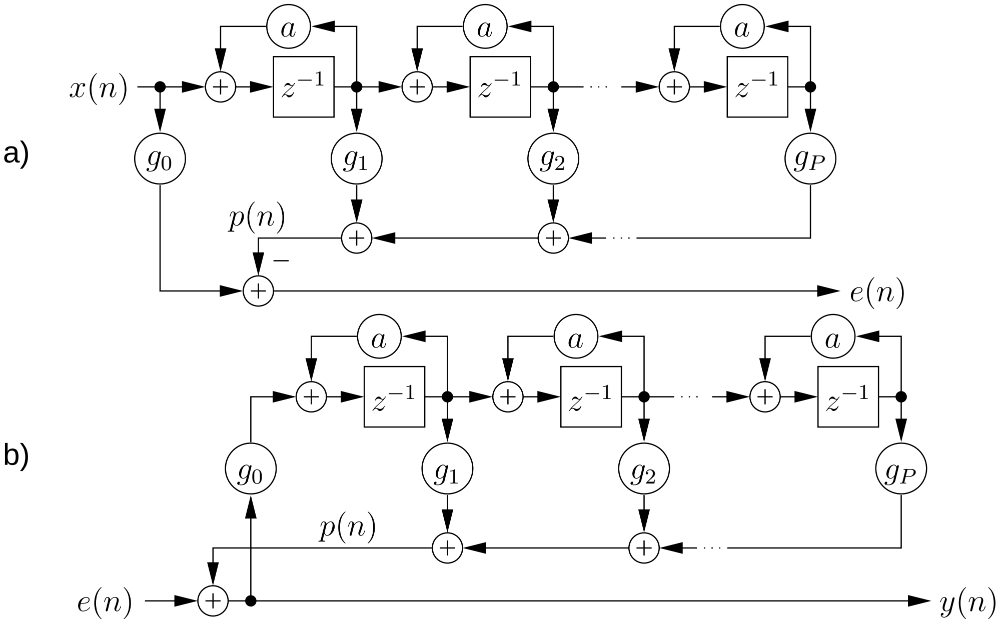

Fig. 6. Structure of a pre-filter (a) and a post-filter (b) with frequency warping.

The advantage of frequency warping for the approximation of masked thresholds can be seen in Figure 7 which shows the short term spectrum of a segment of a speech signal, the masked threshold generated by the psychoacoustic model, and the post-filter frequency responses without warping (a) and with warping (b). A warping coefficient $a = 0.5$ was used in the second case.

The coefficients for a frequency warped pre- and post-filter can be easily derived from a masked threshold with the same procedure as described above for the non-warped case. The only difference is that the samples which are given to the inverse DFT now have to be taken at non-uniformly spaced frequencies according to the non-linear frequency warping function.

4 FILTER ADAPTATION

For the temporal adaptation a psychoacoustic model with sufficient temporal resolution needs to be applied to control the update of filter parameters. First subjective evaluations have shown that an update interval of approximately $2 \ldots 4$ ms is appropriate for achieving a high audio quality. However these experiments also showed that simple switching of filter parameter sets from one time interval to the next can lead to

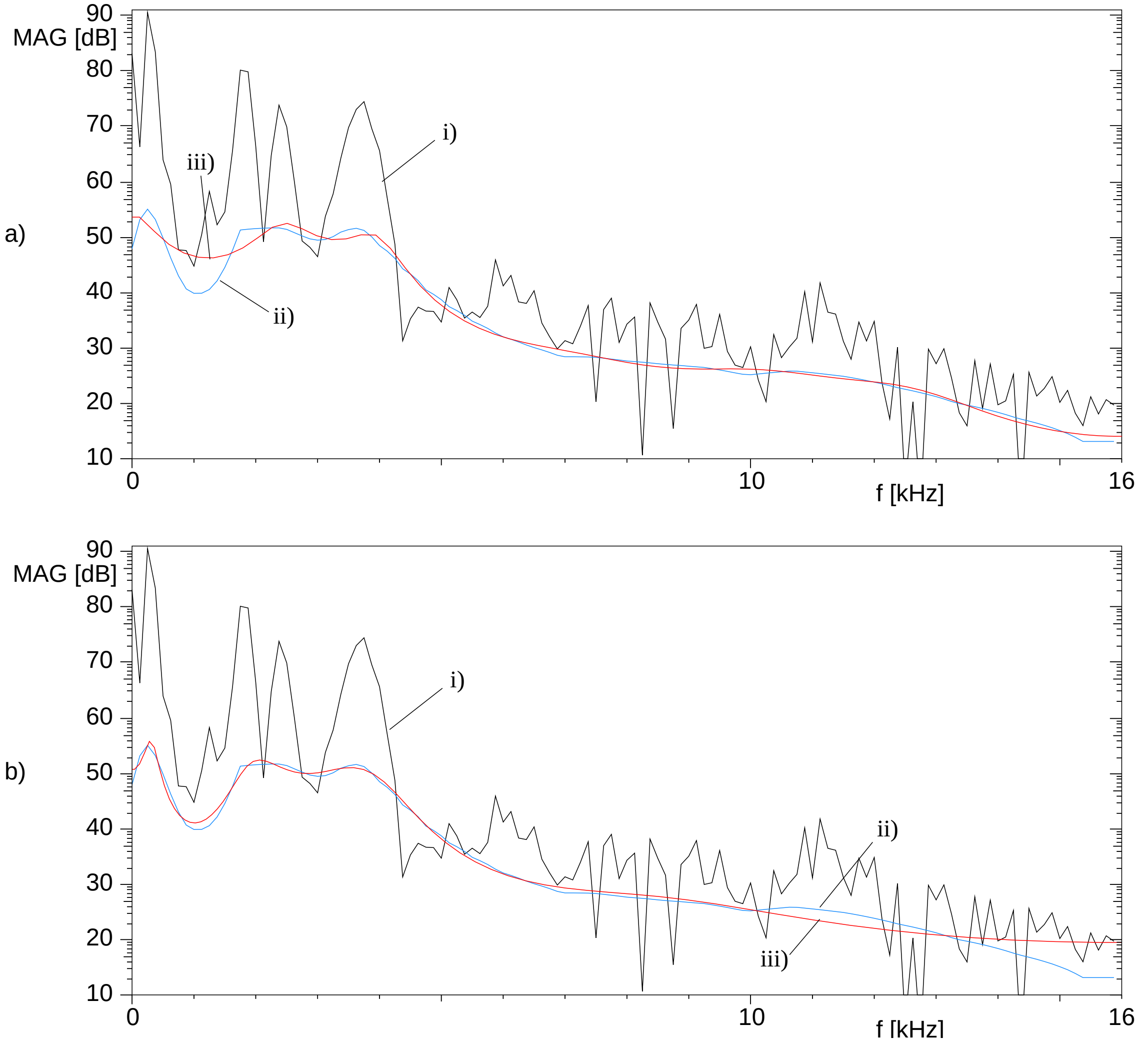

Fig. 7. Short term spectrum (i), masked threshold (ii) and its approximation (iii) without warping (a) and with warping using $a = 0.5$ (b).

audible artifacts. Therefore interpolation techniques for generating filter parameter sets at a higher temporal resolution were considered. The following approaches were investigated and will be described in more detail:

- linear interpolation of filter coefficients,
- Line Spectral Frequency (LSF) interpolation,
- linear interpolation of coefficients in a lattice structure.

4.1 Linear interpolation of filter coefficients

The most obvious approach for avoiding rapid changes of filter coefficients is a direct interpolation. In this case each filter coefficient is linearly interpolated between the values corresponding to two subsequent update instances. However problems can arise from the fact that due to the feedback structure the post-filter is a filter with an infinite impulse response (IIR) which can become instable. Even when interpolating the coefficients between two parameter sets of stable filters, stability is not guaranteed for the interpolated coefficient sets. While an unquantized signal still would be perfectly reconstructed, quantization errors could be amplified in an uncontrolled way. Experiments with this interpolation technique showed that these problems really occurred in the practical application and led to audible artifacts. These effects were even stronger with frequency warped filters, since here the pre-filter is also IIR.

4.2 LSF interpolation

In LPC based speech coders predictor coefficients are often converted to so-called Line Spectral Frequencies, which allow the use of efficient coding techniques [8]. Additionally they have the nice property that linear interpolation between different parameter sets always leads to a stable system. However the complexity of the required conversion operations between filter coefficients and LSF's limits the frequency of updates. Therefore the LSF update intervals were limited to a range of $0.5 \ldots 1$ ms. However, this still led to audible artifacts for some critical signals, even when combined with a sample-wise filter coefficient interpolation.

4.3 Linear interpolation of coefficients in a lattice structure

Another representation for predictor coefficients which is frequently used in LPC systems is based on a lattice filter structure [9]. The parameters which specify the filter characteristic are often referred to as reflection coefficients. These coefficients also have the property that stability is guaranteed for interpolation between parameter sets of stable systems. The other advantage is that they can directly be used in the mentioned lattice filter structure so that no complex conversions are necessary. Frequency warping can also be used for the lattice structure. For this purpose again allpass systems have to replace delay units. The resulting structure of a pre-filter with reflection coefficients $k_1 \ldots k_P$ is shown in Figure 8, where $A(z)$ stands for an allpass system, as shown in Fig. 5. As mentioned, in contrast with delay elements z^{-1} the allpass system contains a non-delayed path from input to output. Hence, due to the feedback in the corresponding post-filter, the problem of a zero-lag loop arises again. This can be seen in Figure 9. Currently there are no lattice structures known which allow the use of frequency warping without introducing a zero-lag loop in the decoder.

Fortunately a strategy for implementing adaptive filters including a zero-lag loop due to feedback was presented recently [10]. It is based on a separate calculation of two different components of the output value and performs the filter operation in three steps. A first part of the output value is calculated which only depends on the current input value while neglecting all outputs of delay elements (i.e. assuming that they are zero). In this step the zero-lag loop has to be considered as a gain factor which can be obtained by solving a linear equation for the overall input/output relation. In the next step the second part of the output value is calculated which only depends on the outputs of the delay elements with the input value assumed to be zero. In this step the zero-lag loop has no effect. The two components are then added to obtain the final output value. This can then be used to calculate the input values of the delay elements, i.e. their next output values.

This lattice filter structure allows a sample-wise interpolation without stability problems and without excessive complexity. The conversion to reflection coefficients only needs to be done at the regular update intervals. Subjective evaluation verified that the transition problems were solved.

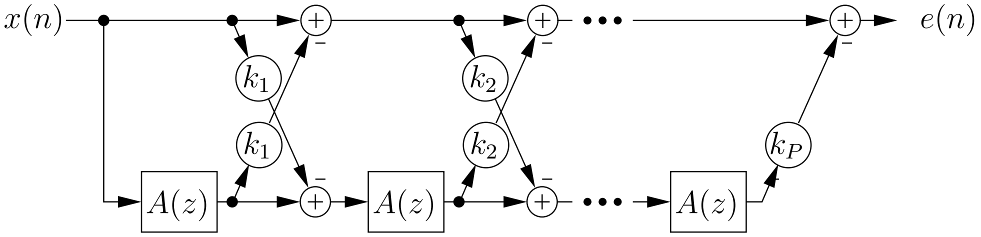

Fig. 8. Lattice structure for pre-filter ($A(z)$: allpass system for frequency warping, as in Fig. 5).

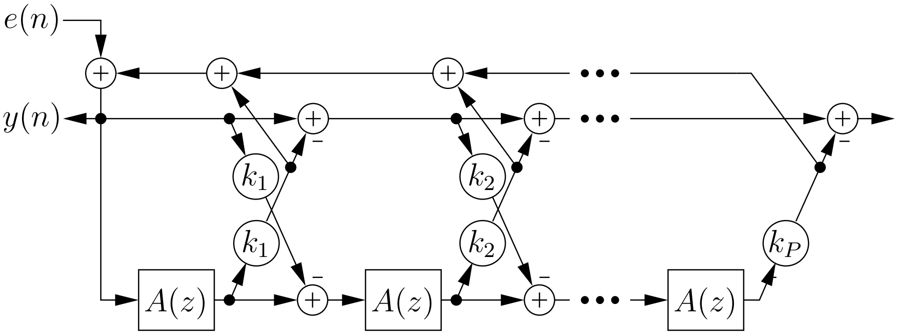

Fig. 9. Lattice structure for post-filter ($A(z)$: allpass system for frequency warping, as in Fig. 5).

5 PARAMETER CODING

Efficient coding of the side information which needs to be transmitted to control the post-filter in the decoder is another important factor for the success of the new technique.

Although the lattice structure was selected for a sample-wise parameter interpolation, the LSF representation proved to be the most suitable for efficient quantization and coding. Due to the fact that masked thresholds are much smoother than short term signal spectra and due to the use of frequency warping, a filter order of $P = 12$ was found to be sufficient even at a sampling rate of 48 kHz. For the LSF quantization and coding a vector quantizer (VQ) as frequently used in speech coding was applied. The codebook was trained on a large set of input samples and a surprisingly low number of 20 bit per vector was found to be sufficient.

Updating the filter coefficients every $2 \ldots 3$ ms and using the described LSF-VQ would result in a side information bit rate of $7 \ldots 10$ kbit/s. This is an acceptable value for high quality coding at a total bit rate above 32 kbit/s, but for very low bit rate coding a further reduction is necessary. Therefore an adaptive selection and interpolation scheme was developed. It allows to select the update rate in dependence of the amount of variations in the masked threshold. It compares the current masked threshold to the last threshold for which filter parameters have been transmitted. A new parameter set is only transmitted, if the difference exceeds a given limit. And even if the limit is exceeded, a further check is performed which can activate an interpolation scheme rather than transmitting individual parameter sets. The control data for the selection of filter parameter sets and for the interpolation mode is transmitted in addition to the selected filter data. Because it is important that the pre-filter and the post-filter use exactly the same filter coefficients, the decoding and interpolation also is performed in the encoder (Figure 10).

In a first implementation this scheme was adapted to the frame length of the transform coder used for transmission of the pre-filter output signal. With a frame length of 1024 samples and a sub-frame length of 128 samples for the parameter updates the number of transmitted parameter sets per frame can vary between 1 and 8. In addition two bits per sub-frame are transmitted to signal the interpolation mode. This results in a side information rate between 1.7 and 8.25 kbit/s at a sampling rate of 48 kHz. If the same scheme is used at lower sampling rates these values are reduced accordingly.

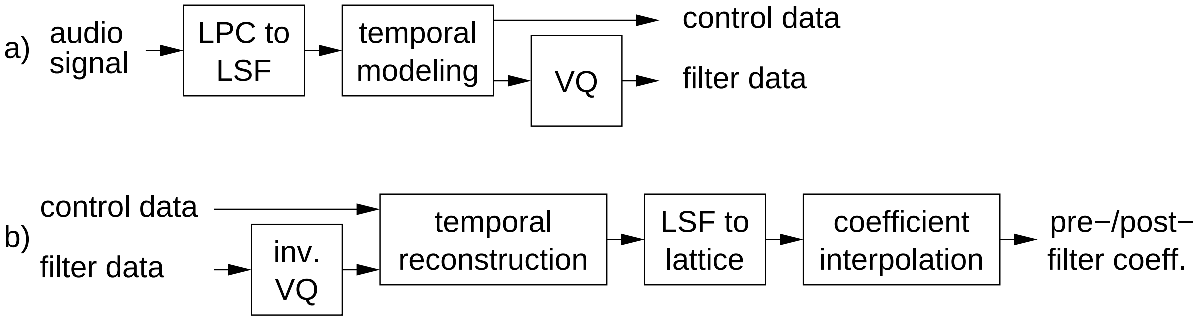

Fig. 10. Adaptation of pre- and post-filter, a) modules only needed in the encoder, b) modules needed in encoder and decoder.

6 RESULTS

In a first implementation the pre- and post-filter technique was integrated in the PAC audio coder [3], such that only its transform coding parts were used. This system was also used for the selection of the best filter structure and for the optimization of its parameters. The outcome was to use the warped lattice structure with sample-wise coefficient interpolation as described above.

In informal listening tests, the resulting system was compared to PAC in its original form and to various codecs, including Liquid Audio AAC, RealSystem G2, Opticom MP3, operating at a bit rate of 16 kbit/s. The test material included speech, speech plus background, and music signals [11]. The items used for the training of the LSF vector quantizer were not included.

The general outcome of these evaluations was that the new system was judged to provide the best overall quality, and it showed the most consistent quality across the different types of audio signals of all systems. The improvements were particularly pronounced for speech signals in comparison with other audio coders.

The same general trend in comparison to the original PAC was observed when bit rates in the range of $24 \ldots 32$ kbit/s were used.

7 CONCLUSIONS

A new perceptual audio coding paradigm was presented, which is based on a separation of irrelevancy reduction and redundancy reduction. This separation allows a better optimization and adaptation to the time-varying masked threshold for exploiting properties of the human auditory system, and to the signal properties for maximizing the coding gain. The irrelevance reduction part consists of a time-varying pre- and post-filter modeling the masked threshold. The redundancy reduction part can be implemented with the efficient modules for transform, quantization, and coding of a transform coder.

To obtain a temporal and spectral resolution similar to that of the human auditory system, frequency warping is applied to the filter structure. To minimize the amount of side information required for filter adaptation, coding and transmission of the filter parameters is based on LSF vector quantization in combination with a flexible selection and interpolation scheme .

At bit rates around 1 bit/sample the resulting system significantly improves the subjective quality especially for signals containing speech. Compared to other currently available coding techniques it seems to provide the best average quality at bit rates around 16 kbit/s.

Further optimization for achieving near-CD quality should focus on the psychoacoustic model which needs to provide a sufficient spectral and temporal resolution.

Future applications of the pre- and post-filter concept even might include coding techniques other than transform coding. One example would be the use of lossless coding techniques in combination with a fixed scalar quantizer.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

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

[5] B. Edler and G. Schuller, “Audio coding using a psychoacoustic pre- and post-filter,” in *Proc. ICASSP-2000*, 2000.

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

[7] U. K. Laine, M. Karjalainen, and T. Altosaar, “Warped linear prediction (WLP) in speech and audio processing,” in *Proc. ICASSP-94*, pp. III–349 – III–352, 1994.

[8] F. K. Soong and B.-H. Juang, “Line spectrum pair (LSP) and speech data compression,” in *Proc. ICASSP-84*, pp. 1.10.1–1.10.4, 1984.

[9] F. Itakura and S. Saito, “Digital filtering techniques for speech analysis and synthesis,” in *7th Int. Congr. Acoustics*, 1971.

[10] A. Härmä, “Implementation of recursive filters having delay free loops,” in *Proc. ICASSP-98*, pp. 1261–1264, 1998.

[11] Institut für Rundfunktechnik, “http://radio.irt.de/aida/demo/28km_e.htm,” *Comparison of Different Internet-Radio Audio Systems*.