WAVEFORM APPROXIMATING RESIDUAL AUDIO CODING WITH PERCEPTUAL PRE- AND POST-FILTERING

Jesper Kjær Nielsen, Jesper Rindom Jensen, Mads Græsbøll Christensen, Søren Holdt Jensen and Torben Larsen

Aalborg University
Department of Electronic Systems
Niels Jernes Vej 12, DK-9220 Aalborg
E-mail: {jkjaer,jesperrj,mgc,shj,tl}@es.aau.dk

ABSTRACT

We investigate waveform approximating residual coding for a sinusoidal parametric audio coder at low bit rates. The residual coding is based on the well-known pre- and post-filtering method with lossless coding [1] which features perceptual weighting for short time segments. We compare the incurred perceptual distortion from joint quantization of the residual and the sinusoids for different bit rates. In addition to that, we develop a transform coding scheme for the coefficients in the pre- and post-filters which must be send as side information between the encoder and decoder. Our investigations show that the combination of the sinusoidal subcoder and the pre- and post-filtering entails an overall lower perceptual distortion for low as well as high bit rates. Also, the developed transform coding scheme enables efficient coding of the side information at a low bit rate.

Index Terms— Perceptual audio coding, residual coding, pre- and post-filtering, sinusoidal audio coding

1. INTRODUCTION

The reduction of the bit rate for a given fidelity in audio coders has been subject to extensive research in the past few decades. This has led to a variety of audio coders of which MPEG-1 layer 3 (MP3) and MPEG-2/4 Advanced Audio Coding (AAC) are the most widespread. These audio coders can typically achieve CD-quality at bit rates of 96 kbit/s and 64 kbit/s for a mono signal [1], respectively, whereas the standard pulse code modulation (PCM) entails a bit rate of 705.1 kbit/s for a mono signal with 16 bit/sample and a sample rate of 44.1 kHz. The large compression factor is achieved by use of perceptual audio coding which comprises irrelevance and redundancy removal. Irrelevance removal is a lossy process in which inaudible signal components are discarded. Inaudibility is determined by a masking curve derived from a psycho-acoustical model, and it depends on the time, frequency and amplitude characteristics of the audio signal [2]. Redundancy removal is a lossless process that removes statistical dependencies within the signal.

Traditional audio coders use subband coding and/or transform coding (see e.g. [3, 4]) in which an audio signal in the encoder is transformed into a perceptual domain where quantization according to a derived masking curve is succeeded by lossless coding. In the decoder, the inverse transform is applied and this produces the reconstructed audio signal. For very low bit rates, however, subband coding and transform coding are not optimal for some audio signals for which reason parametric audio coding has been used as an alternative in the recent years [5]. In parametric audio coding a model of the audio signal is assumed, and the encoding process thus reduces to an estimation of the parameters in the assumed model. A very popular parametric model is the sinusoidal model which recently has been standardized as MPEG-4 HILN (harmonics and individual lines plus noise) [6]. HILN consists like most other parametric coders of a sinusoidal subcoder and a residual or noise subcoder where the latter codes the remaining audio signal which is not extracted by the sinusoidal subcoder. Residual subcoders are typically divided into non-waveform and waveforms approximating coders. The non-waveform approximating subcoders are often based on stochastic modelling of the residual and perform well at low bit rates. The audio quality, however, does not in general increase with increasing bit rate which is in contrast to the waveform approximating coders whose main drawback is poor performance at low bit rates.

In this paper, we investigate a waveform approximating residual subcoder for low bit rates in combination with a simple sinusoidal subcoder. The residual subcoder is based on the pre- and post-filtering method [1, 7] which features perceptual weighting for very short time segments and de-correlation with the weighted cascade least-mean-square (WCLMS) prediction. In the pre- and post-filtering method, a pre-filter adapts its frequency response to the inverse of the masking curve thus mapping the audio signal to a perceptual domain in which irrelevance reduction can be performed in a straight-forward manner. The inverse filtering is performed by the post-filter whose frequency response equals the masking curve. The adaption of the pre- and post-filter to the masking curve requires side information to be send from the encoder to the decoder. In this paper, an efficient encoding scheme is also proposed based on transform coding with the fixed Karhunen-Loeve Transform (KLT).

The paper is organized as follows. In Section 2, we briefly present the pre- and post-filtering method, sinusoidal audio coding and 2-dimensional transform coding. Based on this, our implementation of the sinusoidal subcoder and the pre- and post-filtered residual subcoder is given in Section 3 along with a description of the implementation of the transform coding of the masking curves. In Section 4, the results are presented while Section 5 concludes the paper.

The work of M.G. Christensen was supported by the Parametric Audio Processing project, Danish Research Council for Technology and Production Sciences grant no. 274–06–0521

The work of S.H. Jensen was partly supported by the Danish Technical Research Council, through the framework project Intelligent Sound, www.intelligentsound.org (STVF No. 26-04-0092)

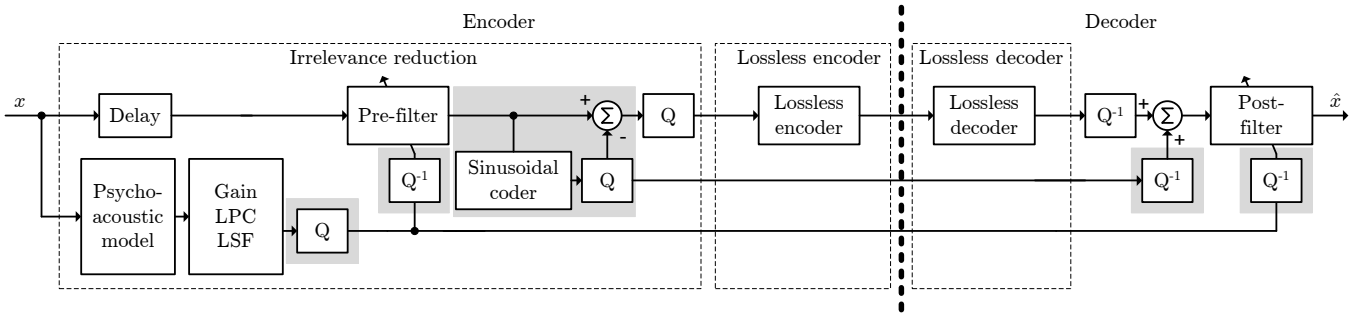

Fig. 1. Block diagram of the sinusoidal subcoder integrated in the pre- and post-filtering setup where the WCLMS predictor acts as a waveform approximating residual subcoder. The shaded backgrounds indicates our modification of the original pre- and post-filtering setup.

2. FUNDAMENTALS

2.1. Pre- and Post-Filtering

The overall pre- and post-filtering system consists of an encoder and a decoder as depicted in Fig. 1. In the encoder, irrelevance and redundancy removal is separated into two different parts. The irrelevance removal is performed by adaptive psycho-acoustical controlled pre- and post-filters, whose frequency responses are determined by the masking curve, and a uniform quantizer. The redundancy removal is performed by a lossless encoder based on weighted cascade least-mean-square (WCLMS) prediction. The decoder consists of a lossless decoder, an inverse quantizer and the post-filter whose frequency response is the inverse of the pre-filter and, hence, equals the masking curve. The masking curve, obtained from the psycho-acoustic model, is parametrized using warped linear predictive coding (WLPC) [8] and the resulting WLPC coefficients and prediction error standard deviation are used in the frequency warped pre- and post-filters as filter coefficients and gain factor, respectively. The masking curves, and thus the WLPC coefficients, are updated every 2 ms to 4 ms. Since a direct switch between old and new filter coefficients leads to audible artifacts [1, 7], interpolation between the coefficients is necessary. However, the interpolation requirement introduces stability issues in the post-filter since WLPC coefficients are not suitable for interpolation. Therefore, the WLPC coefficients are converted into e.g. the line spectral frequency (LSF) coefficient representation [9] or the reflection coefficient (PARCOR) representation [10] which are both amenable to interpolation. The operation of the post-filter requires the frequency response of the pre-filter to be coded and send as side information between the encoder and decoder. In [7] this is done by use of vector quantization of the LSF coefficients which results in bit rates from 7 kbit/s to 10 kbits/s.

2.2. Sinusoidal Audio Coding

In this paper, we consider the following sinusoidal model of order L for a time frame $n = 0, 1, \cdots, N-1$ of an audio signal $x[n]$

$$x[n] = \sum_{l=1}^{L} A_l \cos(\omega_l n + \phi_l) + e[n] \tag{1}$$

where A_l, ω_l and ϕ_l are the amplitude, frequency and phase of the l'th sinusoid, respectively. The difference $e[n]$ between the signal extracted by this model and the actual audio signal is termed the residual. For each time frame, which might overlap the previous, the parameters of Eq. (1) are estimated by use of some suitable estimator as for example the perceptual matching pursuit (PMP) [11] which iteratively seeks to minimize a perceptual norm based on this residual. The sinusoidal model in Eq. (1) is effective for coding stationary tonal signals, but it entails a lot of problems in coding non-stationary and transient signals. For this reason, many extensions to the basic sinusoidal model have been proposed which among others comprise adaptive segmentation [12] and amplitude modulation [13]. In this paper, however, we will use the simple model in Eq. (1). In order to enable efficient transmission of the sinusoidal parameters between encoder and decoder, the sinusoidal parameters have to be quantized. There exists several approaches for this quantization ranging from simple independent uniform quantizers [13] to more refined dependent trellis-coded quantizers [14].

2.3. 2-D Transform Coding

The WLPC representation of the masking curve requires side information to be send from the encoder and decoder. To lower the overall bit rate, it is therefore desirable to reduce the amount of side information as much as possible, and transform coding is widely used for this compression task. In 2-dimensional transform coding, the relationship between the original matrix $\boldsymbol{X}$ of size $P \times M$ and the transform matrix $\boldsymbol{Y}$ of size $P \times M$ is

$$\boldsymbol{Y} = \boldsymbol{T}_v \boldsymbol{X} \boldsymbol{T}_h^H \quad \text{and} \quad \boldsymbol{X} = \boldsymbol{T}_v^H \boldsymbol{Y} \boldsymbol{T}_h \tag{2}$$

where $(\cdot)^H$ denotes the complex transpose, and $\boldsymbol{T}_v$ and $\boldsymbol{T}_h$ are separable orthonormal transform kernels of size $P \times P$ and $M \times M$, respectively. The main motivation behind transform coding is that, for a suitable pair of transform kernels, the quantization of the transform coefficients in $\boldsymbol{Y}$ leads to a smaller overall distortion as compared to direct quantization of $\boldsymbol{X}$ for the same bit rate. It can be shown that the optimum linear transform is the Karhunen-Loeve Transform (KLT) whose transform kernels are found from the eigenvalue decomposition in horizontal and vertical direction, denoted by subscripts h and v, respectively, of $\boldsymbol{X}$ given by

$$\boldsymbol{R}_v = \boldsymbol{T}_v^H \boldsymbol{\Lambda}_v \boldsymbol{T}_v \quad \text{and} \quad \boldsymbol{R}_h = \boldsymbol{T}_h^H \boldsymbol{\Lambda}_h \boldsymbol{T}_h \tag{3}$$

provided that the autocorrelation function of $\boldsymbol{X}$ is separable in horizontal and vertical direction [4]. The main drawback of the KLT is that it depends on the statistics of $\boldsymbol{X}$. For this reason other suboptimal, but fixed, transforms such as the discrete Cosine Transform (DCT) and the discrete Fourier Transform (DFT) have been suggested. Another advantage of these transforms is that they can be implemented in an effective way using the FFT.

For quantization of the transform coefficients, entropy coded scalar quantization is considered since it fits well in the sinusoidal

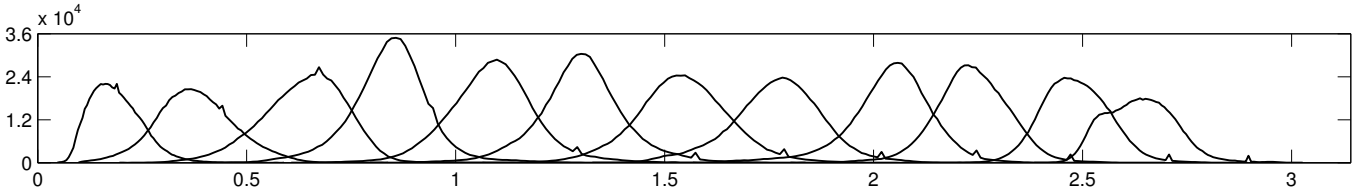

Fig. 2. Histograms with 100 bins for the LSF coefficients. The histograms are computed from the training database consisting of approximately 600,000 LSF vectors.

coder and pre- and post filtering setup. It can be shown analytically that the uniform quantizer under high resolution approximations is the optimal quantizer for entropy coded scalar quantization with an average rate only 0.255 bits from the Shannon lower bound [3]. Empirical studies have also shown that the uniform quantizer is nearly optimal for low bit rates as well [3]. Under high resolution approximations, the optimal step size Δ of all the uniform quantizers are equal and can be shown to be equal to

$$\Delta = 2^{-\bar{R}} \left[\prod_{i=1}^{P} \prod_{j=1}^{M} 2^{h(y_{ij})} \right]^{\frac{1}{PM}} \tag{4}$$

where $\bar{R}$ is the desired average bit rate and $h(y_{ij})$ denotes the differential entropy of the (i,j)'th transform coefficient. The differential entropies for different probability density functions (pdf) can be found in e.g. [4].

3. IMPLEMENTATION

The coding system considered in this paper is based on the simple sinusoidal subcoder implementing the signal model in Eq. (1) and the pre- and post-filter setup acting as a waveform approximating residual subcoder. The pre- and post-filters were implemented using a warped lattice structure specified by PARCOR coefficients, a gain factor and a warping coefficient. To avoid audible artifacts in the coded signals due to rapidly changing filter coefficients, linear interpolation was applied on the PARCOR coefficients and the gain factor for each input sample as in [7]. Fig. 1 depicts the block diagram of the considered coder. The sinusoidal subcoder is placed between the pre- and post-filter in order to utilize the high time resolution of the perceptual weighting in the pre- and post-filtering setup. This also saves the separate perceptual weighting in the sinusoidal subcoder thus reducing the computational complexity of the sinusoidal subcoder from that of perceptual matching pursuit to that of matching pursuit. The cost of this choice is the sub-optimality introduced by the WLPC representation of the frequency response of the pre- and post-filters. With this setup, the residual subcoder is thus constituted by the uniform quantizer and the WCLMS-based lossless coder.

Efficient coding of the LSF coefficients is widely studied in the field of speech coding, but has not been treated in great detail for coding of the masking curve. The studies in speech coding comprise among others the statistical properties of the LSF coefficients [15], vector predictive quantization [16] and transform coding [17] of the LSF coefficients. They show that LSF coefficients are highly correlated in the same frame and between frames and that the distributions of the LSF coefficient resemble skewed Gaussian and Laplace distributions. In our studies, the LSF coefficients describe the prewarped masking curve, but they seem to have the same statistical properties as the LSF coefficients in speech coding. The histograms for the LSF coefficients, each using 100 bins, are shown in Fig. 2. These results were found from an analysis of a music training database consisting of eight different songs of a total length of approximately 40 minutes. An LSF column vector of dimension $P = 12$ was computed for every music frame of 4 ms which led to a training database of approximately 600,000 LSF vectors.

In our coding scheme, transform coding is used to code the LSF coefficients as opposed to vector quantization in [7]. The motivation behind this choice is that transform coding enables the use of simple scalar quantizers. We use and compare the performance of fixed KLT, DCT and PCM where fixed KLT refers to that the transform kernels are found from the training database and fixed during coding. The PCM uses simply the identity matrix as transform kernels. For quantization the entropy coded scalar quantizers are used, and the step size Δ is found from Eq. (4) with the Gaussian distribution as the model for the transform coefficients since it was found to have the greatest resemblance with the estimated distribution of the transform coefficients shown in Fig. 2. The variances of the transform coefficients were found from the training database. Another music database consisting of approximately 84,000 LSF vectors was used for testing.

4. EXPERIMENTAL RESULTS

The following section describes the results obtained from the measurements of the sinusoidal subcoder combined with the WCLMS based residual subcoder. Also, the evaluation of the transform coding of the LSF coefficients is described.

4.1. Sinusoidal Coding with Residual Coding based on Pre- and Post-filtering

The evaluation of the sinusoidal subcoder in combination with the WCLMS based residual coder was performed by means of rate-distortion measurements at different bit rates. In these measurements, the PARCOR coefficients as well as the gain factor were found from WLPC of the masking curve derived from 4 ms non-overlapping time segments of the input signal using the psychoacoustical model in [18]. The WLPC coefficients were transformed to the LSF representation and quantized using transform coding as described in Section 4.2. For the sinusoidal coding, 32 ms 50 % overlapping von Hann windowed time segments were used, and 4 ms time segments for the pre- and post-filtering. Further, the phases of the sinusoids were quantized uniformly using 5 bits while the amplitudes and frequencies were quantized in the logarithmic domain using step sizes of 0.161 and 0.003, respectively, as in [13]. The step size of the uniform quantizer and the number of sinusoids in the sinusoidal subcoder were the only parameters that were varied in the measurements, and a (R, D)-pair was calculated from a (step size, number of sinusoids)-pair. The distortion was measured as the mean-square-error (MSE) in the perceptual domain, i.e. the MSE

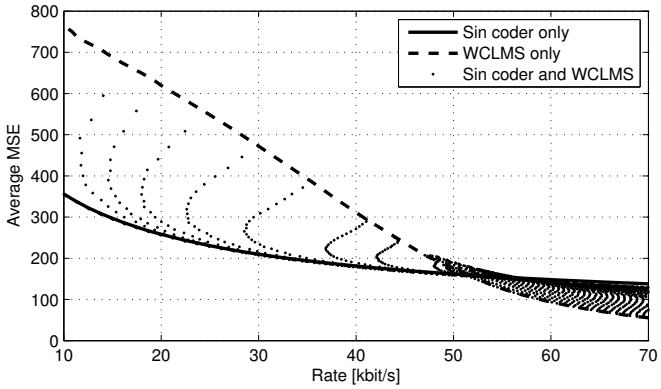

Fig. 3. Measured average rate-distortion for different (step size, number of sinusoids)-pairs for a music piece by Clapton.

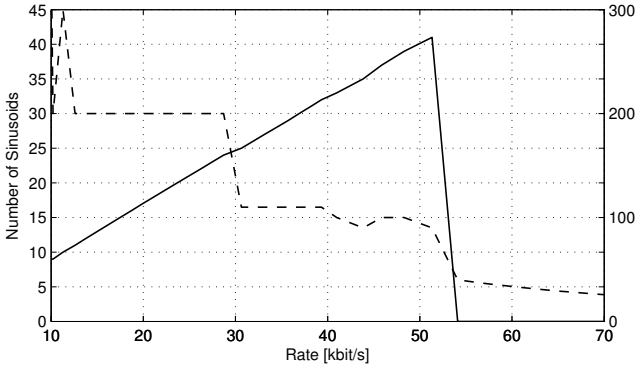

Fig. 4. Optimal number of sinusoids (solid) and optimal step size (dashed) associated with a music piece by Clapton.

between the output signal of the pre-filter and the input signal of the post-filter. The rate was measured as the sum of the estimated output entropy of the WCLMS-based lossless encoder and the estimated entropy of the quantized frequencies, phases and amplitudes of the sinusoids. Thus, the rate did not include the quantized LSF coefficients and gain factor. This contribution is found in Section 4.2 and must be added as well in order to obtain the total rate.

The R-D measurements were performed on two different music pieces: One with a stochastic and transient behaviour without vocal (10 seconds from intro of live recording of Layla by Eric Clapton - Fig. 3 and Fig. 4) and one with a mixture of music and vocal (10 seconds from Head over Heels by Abba - Fig. 5). Fig. 3 and Fig. 5 showed the same trend. For low bit rates, the perceptual MSE was minimized if most of the bits were allocated to the sinusoidal subcoder whereas for high bit rates, the perceptual MSE was minimized if all of the bits were allocated the uniform quantizer in the residual subcoder. The transition region was quite small and occured at a rate of approximately 50 kbit/s. Fig. 4 shows the optimal number of sinusoids and optimal step size associated with Fig. 3. It shows that the number of sinusoids was increasing linearly until some point over which the number of sinusoids was decreased to zero almost immediately. At the same point the quantizer step size was decreased dramatically.

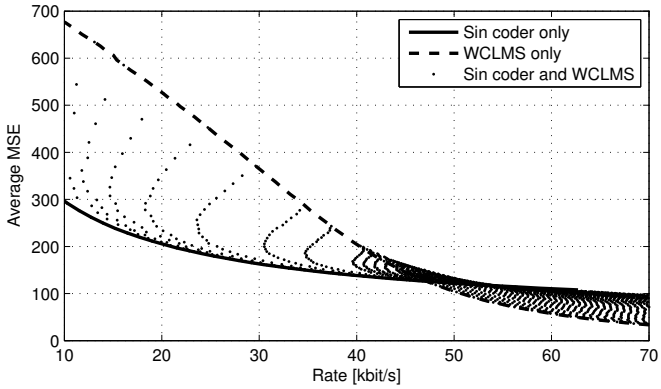

Fig. 5. Measured average rate-distortion for different (step size, number of sinusoids)-pairs for a music piece by Abba.

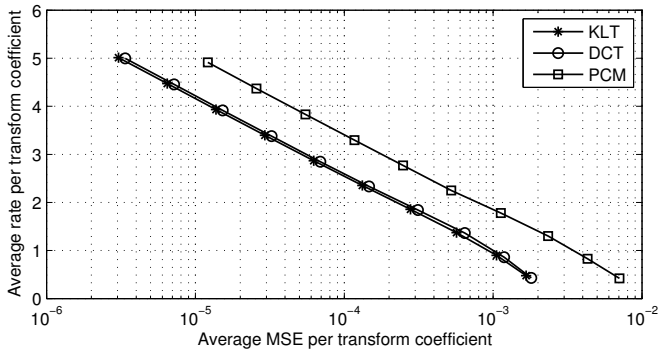

Fig. 6. Measured average rate-distortion pairs for each transform coefficient of the fixed KLT, the DCT and PCM. The measurements were performed on a test database of approximately 84,000 LSF vectors.

4.2. Transform Coding of Masking Curves

The evaluation of the transform coding scheme of the masking curves was performed by use of two tests based on the LSF vectors in the test database: 1) Rate-distortion measurements (R-D) for the fixed KLT, the DCT and PCM and 2) log spectral distortion measurements for the fixed KLT, the DCT and PCM. Fig. 6 depicts the measured R-D points for the fixed KLT, the DCT and PCM operating on a block of $M = 10$ consecutive LSF column vectors of size $P = 12$. The distortion was measured as the mean squared error (MSE) between the unquantized and quantized transform coefficient, and the average rate was estimated as the average of the entropy of each transform coefficient. Fig. 6 shows that the fixed KLT was slightly better than the DCT while PCM performed much worse than both the fixed KLT and the DCT. Since the DCT enables the use of a fast implementation by means of the FFT, the DCT may therefore be the best choice for many application in which computational complexity matters.

Since the MSE distortion measure does not in general correspond to subjective measures, the system performance was also evaluated using the log spectral distortion (LSD) measure which is often used for evaluation of speech coders [17]. The LSD measures the average mean square logarithmic distance between the original and

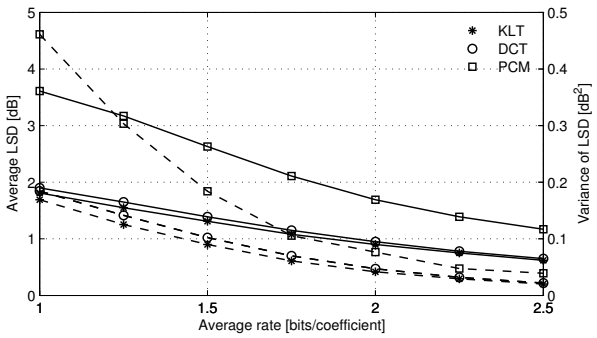

Fig. 7. Mean (solid) and variance (dashed) of LSD for the LSF vectors in the test database for different measured average bit rates.

reconstructed power spectral density (PSD) and is defined as [19]

$$D_{\mathrm{LS}} = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[10\log_{10}\frac{S(\omega)}{\hat{S}(\omega)}\right]^2 d\omega} \qquad (5)$$

where $S(\omega)$ and $\hat{S}(\omega)$ are the original and reconstructed PSD, respectively, which in our setup correspond to the original and reconstructed masking curve. The LSD was computed for each LSF vector in our test database and the sample mean and sample variance for the fixed KLT, the DCT and PCM for different bit rates were calculated. Fig. 7 shows a plot of the measured values. Clearly, the mean value and variance of the fixed KLT and the DCT were significantly lower than that of PCM which resembled the observed pattern of the objective distortion measure in Fig. 6. That is, the fixed KLT was slightly better than the DCT and much better than PCM. It is interesting to note that an average LSD of 1 dB for the fixed KLT resulted in an average bit rate of approximately 1.8 bits per transform coefficient. With a frame length of 4 ms and $P = 12$ this amounts to a bit rate of 5.4 kbit/s which is lower than 7 kbit/s to 10 kbit/s obtained in [7] by use of vector quantization. In speech coding, an LSD of 1 dB is typically considered as a limit of perceptual significance [17].

5. CONCLUSION

In this paper we have focused on two topics. First, we investigated the combination of a simple sinusoidal subcoder and a waveform approximating residual subcoder for low bit rates based on pre- and post-filtering. The results showed that it was possible to use the sinusoidal subcoder with the chosen residual subcoder to reduce the perceptual distortion at low bit rates, whereas, for higher bit rates, the perceptual distortion was lowest using only a WCLMS based subcoder, which is a part of the chosen residual subcoder system. In order to keep the distortion as low as possible for all bit rates, the two coding structures should be allocated bits jointly in and rate-distortion optimal way. Second, it was investigated how the WLPC coefficients for the pre- and post-filters could be coded in an efficient way. It was shown that by applying transform coding using the fixed KLT and by applying entropy coded scalar quantization of the transform coefficients, the MSE as well as the LSD could be improved compared to using the DCT and PCM. For a block length of 4 ms and filter order of 12 an average LSD of 1 dB could be obtained at a bit rate of 5.4 kbit/s.

6. REFERENCES

[1] G.D.T. Schuller, Bin Yu, Dawei Huang, and B. Edler, "Perceptual audio coding using adaptive pre- and post-filters and lossless compression," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 379–390, Sep 2002.

[2] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, Springer, 2. edition, 1999.

[3] R. M. Gray and A. Gersho, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1. edition, 1992.

[4] N.S. Jayant and P. Noll, *Digital Coding of Waveforms*, Pentice-Hall, Inc., 1. edition, 1984.

[5] B. Edler and H. Purnhagen, "Parametric audio coding," *Proc. Conf. Signal Process.*, vol. 1, pp. 21–24 vol.1, 2000.

[6] H. Purnhagen and N. Meine, "HILN-the MPEG-4 parametric audio coding tools," *Proc. IEEE Int. Symp. Circuits and Systems*, vol. 3, pp. 201–204 vol.3, 2000.

[7] B. Edler and G. Schuller, "Audio coding using a psychoacoustic pre- and post-filter," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. II881–II884 vol.2, 2000.

[8] H. W. Strube, "Linear prediction on a warped frequency scale," *J. Acoust. Soc. Am.*, vol. 68, pp. 1071–1076, 1980.

[9] F. Itakura, "Line spectral representation of linear predictive coefficients of speech signals," *J. Acoust. Soc. Am.*, vol. Vol. 57, No. 1, June 1975.

[10] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall Inc, 1. edition, 1978.

[11] M.G. Christensen and S.H. Jensen, "On perceptual distortion minimization and nonlinear least-squares frequency estimation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 14, no. 1, pp. 99–109, Jan. 2006.

[12] P. Prandoni, M. Goodwin, and M. Vetterli, "Optimal time segmentation for signal modeling and compression," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 3, pp. 2029–2032 vol.3, Apr 1997.

[13] M.G. Christensen and S. van de Par, "Efficient parametric coding of transients," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 14, no. 4, pp. 1340–1351, July 2006.

[14] M.H. Larsen, M.G. Christensen, and S.H. Jensen, "Variable dimension trellis-coded quantization of sinusoidal parameters," *IEEE Signal Process. Lett.*, vol. 15, pp. 17–20, 2008.

[15] F. Soong and B. Juang, "Line spectrum pair (LSP) and speech data compression," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 9, pp. 37–40, Mar 1984.

[16] Y. Shoham, "Vector predictive quantization of the spectral parameters for low rate speech coding," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. Vol. 12, April 1987.

[17] N. Farvardin and R. Laroia, "Efficient encoding of speech LSP parameters using the discrete cosine transform," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1989.

[18] S. van de Par, A. Kohlrausch, G. Charestan, and R. Heusdens, "A new psychoacoustical masking model for audio coding applications," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 1805–1808, 2002.

[19] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. Vol. 24, No. 5, October 1976.