A Switched DPCM/Subband Coder for Pre-echo Reduction

Satheesh S and T.V. Sreenivas

Department of Electrical Communication Engineering
Indian Institute of Science, Bangalore, India
tvsree@ece.iisc.ernet.in

Abstract

Recently, adaptive subband coders based on wavelet packet decomposition and psychoacoustic modelling have been proposed to achieve transparent quality compression of audio signals [1],[2]. While these coders perform well for stationary signals, there is no special mechanism in the coder to prevent the pre-echo artifact when transient signals are encoded. In this paper, we propose a switched DPCM/Subband structure to remove the pre-echo problem. This is achieved through a novel temporally varying bit allocation scheme which is based on the temporal masking properties of the human auditory system. The proposed coder/decoder output is found to be free from the pre-echo artifact even at a lower bitrate than the adaptive subband coder.

1. Introduction

Pre-echo is an artifact arising from the lack of temporal resolution in a transform coder or subband coder when the input signal has sharp transients. The quantization noise spreads across the entire transform window length in the case of a transform coder and for a subband coder, the temporal spreading of noise is governed by the effective length of the impulse response of the synthesis filters [3]. The pre-masking effect is only for a short duration of around 5 ms which is insufficient to mask the quantization noise although in the frequency domain the quantization noise is below the masked threshold. This 'unmasking' of quantization noise in the time domain is heard as pre-echo. Several techniques, such as the bit reservoir, adpative window switching, gain modification and temporal noise shaping are employed in the currently available coders to remove pre-echo [4]. The reason for the poor performance of coders for transient signals is the use of simultaneous masking model even for a frame containing a transient signal. It is inappropriate to use the simultaneous masking model for such a frame because the masker(signal) and the target (quantization noise) are not simultaneously present. The audibility of quantization noise in a transient frame is decided by temporal or non-simultaneous masking properties because, by its very nature, a transient frame has a considerable time where the lack of sufficient masker power can lead to audibility of quantization noise. This paper addresses the two issues of noise spreading and an appropriate psychoacoustic model by : (i) Switching to a time domain Differential Pulse Code Modulation (DPCM) coder for encoding a transient frame so that there is no temporal spreading of noise. (ii) Use of temporal masking thresholds derived from psychoacoustic experiments to allocate bits to the prediction residual.

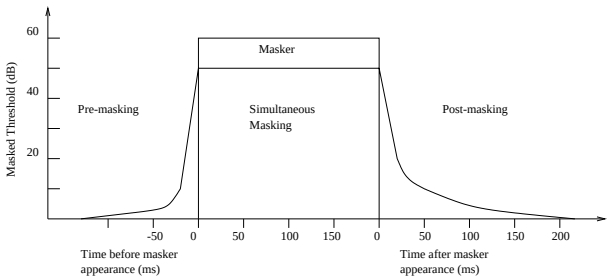

Figure 1: *Temporal Masking.*

2. Temporal Masking Threshold

Temporal or non-simultaneous masking refers to the increase in audible threshold of a target signal occuring before the onset or after the cessation of a masker signal. If the target is present prior to the onset of the masker, the resultant masking is called *Pre-masking or backward masking*. If the target is present after the cessation of the masker, the resultant masking is called post masking or forward masking [5]. The phenomenon of temporal masking is illustrated in Fig. 1. As shown in the figure, pre-masking lasts only for a short duration of about 5ms while post masking extends to 50-300ms. Typical psychoacoustic experiments are performed with pure tones or wideband or narrowband noise as the maskers and targets. However, for audio coding applications, it would be desirable to use more complex stimuli as the maskers and quantization noise as the target, so that the results of the experiment can be easily incorporated into the coder. In this experiment, we use the castanet signal as the masker. The castanet signal shown in Fig. 2 has a strong transient and hence has a potential pre-echo problem due to compression. A frame of length nearly 0.73 seconds containing one transient(spike) is extracted from the castanets signal sampled at 48 KHz. A signal is formed by concatenating multiple instances of this frame separated by silence. The separation between any two spikes is 0.78 seconds and the total duration of the test signal is 4 seconds. The test signal is depicted in Fig. 2. These parameters are fixed so that the silence between the spikes and the total length of the signal is sufficient to make a subjective judgement of quality. Each frame containing a transient is divided into 16 segments, each of 256 samples(= 5.3 ms). The framing is done in such a way that the transient occupies the last segment. For each trial a controlled amount of noise is added to one of the sixteen segments of each transient frame. A typical configuration is shown in Fig. 3. The Signal to Noise Ratio (SNR) is varied from 14 dB to about 50 dB for each segment. The psychoacoustic experiment is performed on three subjects with normal hearing abilities. The sound levels for all

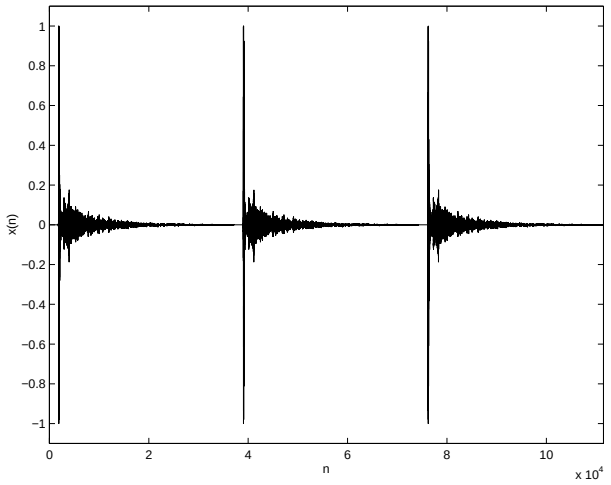

Figure 2: *Masker signal.*

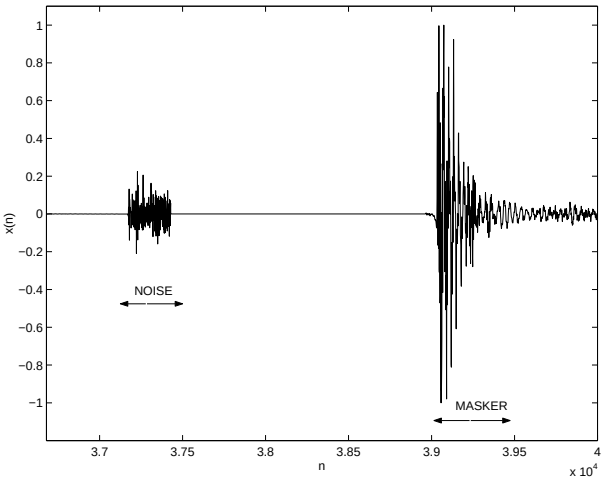

Figure 3: *Noise + Masker signal.*

three subjects are maintained the same. The subjects are first presented the original sample and then the noisy sample after a short pause of 5 seconds. The presentation is using headphones. The subject is asked to rate the quality of the noisy sample in comparison to the original sample on a scale of 1 to 4. The scale is explained in Table 1. The minimum SNR at which the subject is just unable to distinguish between the original and noisy sample is chosen as the threshold SNR for transparent coding. The results of the experiment are shown in Fig. 4. The delay in milliseconds between the cessation of the noise and the onset of the masker is plotted along the x-axis and the threshold SNR is plotted along the y-axis. The graph is in agreement with the temporal masking curve shown in Fig. 1. As the delay between the noise cessation and masker onset increases, the effect of pre-masking decreases and consequently, the threshold SNR requirement for transparency increases. The negative value of delay indicates that the noise cessation occurs later in time than the signal onset. Hence, for this case, the threshold obtained is actually the simultaneous masking threshold. For each value of delay, the average of the threshold SNRs across the 3 subjects is chosen as the threshold SNR for bit allocation in the encoder. If the Signal to quantization ratio of the coded signal is greater than or equal to these threshold values, then perceptual transparency can be achieved.

Table 1: *Subjective scores and their interpretation.*

Score	Interpretation
1	Large quality difference
2	Small quality difference
3	Difficult to make out any difference
4	No difference at all

3. Switched DPCM/Subband coder

The subband coder structure we use is similar to adaptive wavelet packet decomposition. The prototype low pass filter used is a Daubechies20 wavelet basis. The filter bank structure is decided on a frame by frame basis with a view to achieve minimum possible Perceptual Entropy (PE). The input signal is partitioned into frames and the analysis/synthesis filterbank is

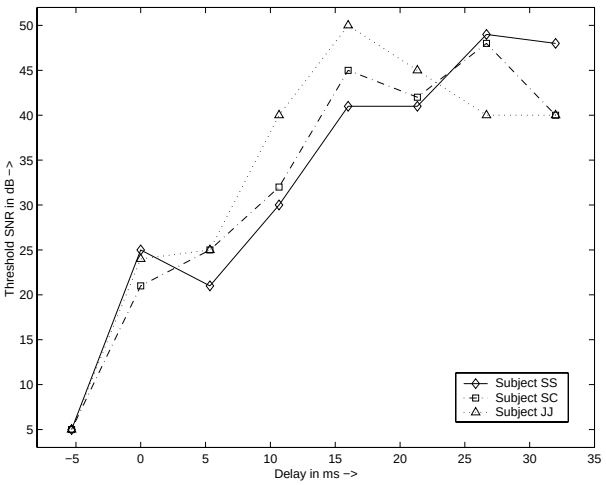

Figure 4: *Threshold SNR v/s delay.*

applied on each frame separately. Since each frame is treated separately, it is necessary to use techniques such as circular extension to ensure critical sampling and avoid boundary effects [6]. Each input frame is 2048 samples long. Depending on the input frame, the filter bank structure can vary from a certain maximum depth of decomposition to no decomposition at all. Though the impulse response of the prototype filter is short, the effective impulse response of the synthesis filters can be long enough to cause pre-echo for large depth of decomposition. The psyhcoacoustic model used here is an implementation of the psychoacoustic model specified in the ISO/IEC standard for MPEG-2 Advanced Audio Codec (AAC). [7]. The AAC psychoacoustic model outputs the masked threshold for each scale factor band. The minimum masked threshold of the scale factor bands spanned by a particular subband is chosen as the masked threshold for that subband. From this masked threshold, the minimum number of bits required to encode the signal transparently is computed for each subband. This quantity is referred to as the subband perceptual entropy. Further binary splitting of a node into two child nodes is carried out if the fol-

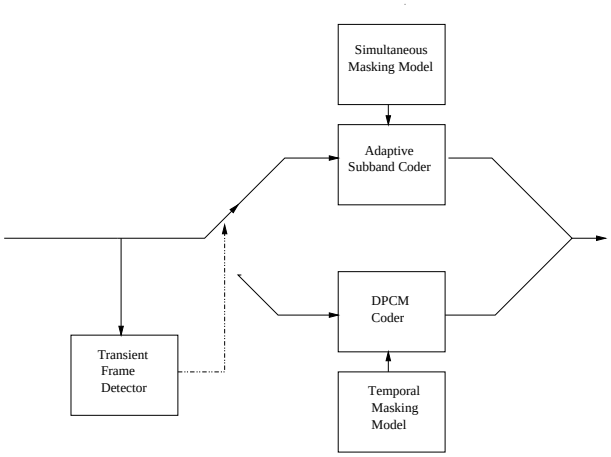

Figure 5: *Block diagram of encoder.*

lowing criterion is satisfied:

$$SPE_{parent} > SPE_{child1} + SPE_{child2} \tag{1}$$

where SPE_{parent} is the subband perceptual entropy of the parent node, SPE_{child1} and SPE_{child2} are the subband perceptual entropies of the two child nodes. That is, a subband is decomposed further only if the decomposition results in a reduction in the perceptual entropy and thereby a reduction in the total bitrate. This approach is similar to that used in [1]. This encoder results in a variable bit rate coder. Although this scheme results in a filterbank structure that minimizes the PE, when the input frame has transients, pre-echo is caused because of the temporal spreading of quantization noise by the synthesis filters and by the use of the inappropriate simultaneous masking model. These issues are addressed by switching to a DPCM based coder which performs bit allocation using a temporal masking model when the input frame has a transient. The block diagram of the resultant coder is shown in Fig. 5.

3.1. Transient frame detector

To perform the switching between the DPCM coder and the adaptive subband coder, it is necessary to identify input frames having transients. This task is performed by the transient frame detector block. The procedure for transient frame detection is explained as follows. The Hilbert envelope of the input frame is computed [4]. The frame is divided into 8 subframes. The mean value of the Hilbert envelope in each subframe is computed. The ratio of the maximum value of the mean Hilbert envelope to the minimum value is computed. If this ratio exceeds a certain threshold, then the frame is declared to contain a transient. If this ratio is less than the threshold, the frame is assumed to be of stationary nature.

3.2. DPCM coder

The time domain DPCM coder is activated when the transient frame detector detects a transient at the input of the audio coder. A closed loop DPCM coder ensures that the reconstruction error at any instant of time is equal to the quantization error at that instant, or in other words, there is no temporal spreading of quantization noise . The procedure adopted in the DPCM coder is described as follows

*Step 1:*The optimum predictor coefficients are calculated from the input frame of 2048 samples. The order of prediction used is 10. The quantization of the predictor coefficients is carried out in the Line Spectral Frequency(LSF) domain. The LSFs are vector quantized using a 512 word codebook [8]. The distance measure used for the LSF vector quantizer is

$$d(f, f') = \sum_{i=1}^{10} [c_i (f_i - f_i')]^2 \tag{2}$$

where f_i and f_i' are the i^{th} LSFs in the test and reference vector, respectively and c_i are the fixed weights assigned to the i^{th} LSF. The fixed weights are given by

$$c_i = \begin{cases} 1.0, & \text{for } 1 \leq i \leq 8; \\ 0.8, & \text{for } i = 9; \\ 0.4, & \text{for } i = 10. \end{cases} \tag{3}$$

The last two LSFs are assigned lower weights than the rest of the LSFs, which takes into account the better frequency resolution of the human auditory system at lower frequencies. The quantized LSF vector is transmitted to the decoder. This is also used in the encoder to perform prediction.

*Step 2:*Prediction is first performed without quantization to find out the dynamic range of the prediction residual. The frame is divided into 8 subframes of 256 samples each. A separate quantizer will be designed for each of the subframes using the temporal masking thresholds determined by the temporal masking model. To design the quantizers, the maximum value of the unquantized prediction residual d_{max} is obtained for each subframe.

*Step 3:*Now, the temporal threshold SNRs for the current frame are computed based on the temporal masking data derived from the psychacoustic experiments. The threshold data is available as a function of the delay between the cessation of the noise and onset of the masker. Let $Th(\delta)$ denote the temporal threshold SNR corresponding to a delay δ. The masker onset corresponds with the position of the spike in the frame. As in the psychoacoustic experiment, we divide the frame into eight subframes of 256 samples each. The subframe which shows the maximum value for the mean Hilbert envelope contains the transient. This subframe is identified as the transient subframe denoted by t. For each subframe j, the delay $\delta(j)$ with respect to the transient subframe is computed.

$$\delta(j) = (t - j)T \tag{4}$$

where T is the time duration of one subframe. The temporal threshold SNR for the j^{th} subframe is given by

$$T_j = \begin{cases} Th(\delta(j)), & \text{for } \delta(j) \leq 0; \\ Th(0), & \text{for } \delta(j) < 0. \end{cases} \tag{5}$$

*Step 4:*A temporally varying bit allocation is implemented by employing different quantizers for the different subframes. To start with, all subframe quantizers are assigned 1 bit each.

*Step 5:*The prediction is now performed with different quantizers for the different subframes. The Signal to Quantization Noise Ratio (SQNR) is computed for each subframe as follows

$$SQNR_j = \frac{\sigma_x^2}{\sigma_q^2(j)} \tag{6}$$

where σ_x^2 is the variance of the entire frame of input and $\sigma_q^2(j)$ is the variance of the quantization noise in the j^{th} subframe. This method of calculation of SQNR is in agreement with the stimuli used for the temporal masking experiment.

Step 6: For each subframe, the $SQNR_j$ obtained is compared with the temporal masking threshold T_j. If $SQNR_j$ is less than T_j, then one more bit is added to the quantizer of the j^{th} subframe. Otherwise, the quantizer of the j^{th} subframe is frozen.

*Step 7:*If for any subframe j, $SQNR_j$ is less than T_j, then go to *Step 5*. If for all subframes $SQNR_j$ is greater than or equal to T_j, the quantization process is over.

At the end of the above process, we have a quantized signal, with different number of bits allocated for different temporal segments and the SQNRs meeting the temporal threshold SNRs in each subframe.

The information transmitted to the decoder comprises of the following:

- Since the DPCM coder is used as a part of the switched DPCM/adaptive subband coder, 1 bit is required to indicate to the decoder whether DPCM is used or subband coder is used.
- The vector quantized LSF coefficients. 9 bits are needed for this.
- The quantized prediction residuals. The number of bits needed for this depends on the input frame and the temporal masking model.
- Side information about the subframe quantizers. This comprises of the maximum absolute value of the quantizer and the number of bits per sample of the quantizer. The maximum absolute value of the quantizer requires 15 bits per quantizer and the number of bits per sample is encoded using 2 bits per quantizer. For 8 subframes, this information amounts to 136 bits per frame.

4. Results and Discussion

The performance of the switched DPCM/subband coder is evaluated using signals with potential pre-echo problem such as castanets and glockenspiel. The performance of the switched DPCM/subband coder is compared with that of the adaptive subband coder. Fig. 6 (a), (b) and (c) show the initial portion of a frame preceding a transient for the original signal, output of switched DPCM/subband coder and the adaptive subband coder respectively. It is clear that the spreading of noise is practically non-existent for the DPCM output coded at an average bitrate of 2.26 bits per sample while it is present for the subband coder output coded at 2.85 bits per sample. Note that these bitrates mentioned here are without any lossless coding being applied. The SQNR for the adaptive subband coder/decoder output is 17.36 dB while that for the switched DPCM/subband coder output is 20.48 dB. In the subjective listening test, each subject is presented with a pair of signals separated by a short pause of 5 seconds. The first signal is the original signal and the second is the output of the switched DPCM/subband coder or the adaptive subband coder. It is found that the subjects judge the switched DPCM/subband coder's output quality to be better than that of the subband coder's output quality.

5. Conclusion

A switched DPCM/subband coder structure is introduced for transparent encoding of transient signals. A psychoacoustic experiment is conducted to measure the temporal masking thresholds. The temporal thresholds obtained from the experiment are used to perform a temporally varying bit allocation in the DPCM coder. The output of the switched coder is found to be free from the pre-echo artifact. Subjective listening tests show that the pre-echo artifact which is present in the adaptive subband coder is effectively reduced in the switched DPCM/subband coder.

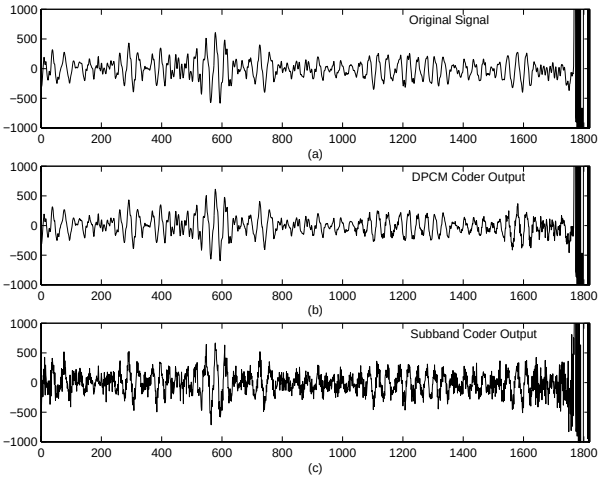

Figure 6: *Comparison of coder outputs* .

6. References

[1] P. Srinivasan and L. H. Jameison, "High quality audio compression using an adaptive wavelet decomposition and psychoacoustic modeling", IEEE Trans. Signal Processing, vol.46,pp.1085-1093, Apr. 1998.

[2] Y. Karelic and D. Malah, "Compression of high quality audio signals using adaptive filterbanks and a zero-tree coder",
Proceedings of Eighteenth Convention of Electrical and Electronics Engineers in Israel, pp 3.2.4-1-3.2.4-5, 1995.

[3] M. Link, "An attack processing of audio signals for optimizing the temporal characteristics of a low bit-rate audio coding system", 95th AES convention, New York 1993, Preprint 3696.

[4] J. Herre and J. D. Johnston, "Enhancing the performance of perceptual audio coders by using temporal noise shaping", 101st AES convention, Los Angeles 1996, Preprint 4384.

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

[6] R. M. Rao and A. S. Bopardikar, Wavelet Transforms Introduction to Theory and Applications, Addison Wesley Longman Inc., 1998.

[7] ISO/IEC, 13818-7, "Information technology-Generic coding of moving pictures and associated audio-Part 7: Advanced audio coding", 1997.

[8] K. K. Paliwal and W. B. Kleijn, "Quantization of LPC Parameters", in W. B. Kleijn and K. K. Paliwal, editors: "Speech Coding and Synthesis", Elseiver, Amsterdam, 1998.