ON THE UTILIZATION OF OVERSHOOT EFFECTS IN LOW-DELAY AUDIO CODING

Aki Härmä, Unto K. Laine, and Matti Karjalainen

Helsinki University of Technology
Laboratory of Acoustics and Audio Signal Processing
P. O. Box 3000, 02015, Espoo, Finland
Aki.Harma@hut.fi

ABSTRACT

In low-delay audio coding (coding delay < 5 ms) there is no time for detailed spectral modeling in the case of brief percussive sounds, e.g., the castanets, and onsets of music or speech sounds. On the other hand, it is known from psychoacoustic experiments that the ear is not accurate near the onset of a wideband sound. In this paper, we study the audibility of coding errors near the onsets of musical sounds in a simulated low-delay audio codec based on frequency-warped linear prediction. It is suggested that for many musical transients it is sufficient to reproduce a rough temporal and spectral envelope of the original signal during the first 5-10 ms. Preliminary listening tests support this idea. It is proposed that the overshoot effect of hearing could be utilized efficiently in enhancing the performance of a low-delay audio coding scheme.

1. INTRODUCTION

Perceptual coding techniques are usually based on subband decomposition and a psychoacoustic model which controls the quantization process within each subband according to a simplified frequency-domain model of hearing. The current authors have studied coding techniques where the auditory model is integrated into the coding process so that no separate models are needed [9]. This has been achieved using so called frequency-warped signal processing techniques where the auditory frequency resolution and frequency masking characteristics of hearing are automatically utilized in the coding process, i.e., the codec may be considered as a coding auditory model. One of the advantages of this approach is that it makes it possible to design a perceptual audio codec having very low coding delay because no additional frequency-domain processing is required.

In [10], it was shown that using a formulation of backward adaptive frequency warped linear prediction it is possible to design a perceptual audio codec for which the coding delay is exactly one sample period. The codec works reasonably well for most of the smoothly varying natural signals, e.g., Suzanne Vega's *Tom's Diner*. The codec is transparent for many natural CD-grade music signals at a bit-rate of 220 kbits/s for stereo. However, it fails completely if a signal contains sudden, intensive transients, e.g., in *Castanets* sequence used in testing by the MPEG community. This is due to the fact that the spectral modeling in the coding process is completely based on past signal values and, hence, the codec has no means to adapt to a sudden previously unknown transient. Therefore, the future versions of the codec are going to be based on a combination of backward and forward adaptive coding and accordingly there will be a small coding delay.

In the speech coding community a low-delay coding usually means that the coding delay should be less than 5 ms, e.g, 2 ms in ITU G.728 [2]. It is reasonable to adopt this rule also to low-delay wideband audio coding. An acoustic signal propagates in air less than one meter within 2 ms and therefore this low coding delay should not produce significant audible echo problems for example in high quality teleconferencing or in applications based on shared virtual acoustics.

On the other hand, 2 ms is 88 samples of an audio signal at 44.1 kHz sampling rate and it does not make possible to estimate a detailed auditory spectral model for an onset of a musical signal within this interval. Accordingly, the performance of a low-delay coding scheme evidently fails in the case of a sudden onset compared to a steady-state situation where it is possible to use past signal values. Nevertheless, it is known from several different listening experiments reported in psychoacoustics literature that the ear is not particularly accurate immediately after the onset of a wideband masker. The purpose on this paper is to study the applicability, potentials and limitations of exploitation of so called simultaneous temporal masking effects in low-delay audio coding.

Section 2 reviews some experimental results from psychoacoustics and discusses their relevance in the coding of transients in music and speech signals. In Section 3, a simulated low-delay coding scheme is introduced and it is verified in listening tests that a significant amount of quantization noise may be accepted during sudden onsets of typical audio and speech signals.

2. PSYCHOACOUSTICS OF TRANSIENTS

It is well understood that the onset of a musical transient is very important for the discrimination of different instruments [12]. However, this applies to the first 50 to 80 ms after the onset which is a relatively long period in low-delay coding.

In this paper, it is suggested that for perceptually transparent reconstruction of various natural signals it is sufficient to reproduce the onset so that it merely has a roughly similar energy distribution in time and frequency during the *first few milliseconds*. This suggestion relies both on a common intuition and some psychoacoustic experiments related to the performance of the ear in those situations.

In the *overshoot* effect the masked threshold increases immediately after the onset of a masker and then decreases rapidly to the level which is obtained in the case of a continuous masker. The effect has been studied extensively in various masker-test signal combinations, see e.g. [17]. For a wideband noise masker and sinusoidal test probes the threshold for a probe is 10-20 dB above the

threshold for continuous maskers during the first 5 to 10 milliseconds after the onset of the masker [3]. The phenomenon disappears completely within the first 100-200 ms. The effect varies significantly as a function of level, bandwidth, individual, etc. Generally, the effect is highest for broadband noise and practically disappears for single tone maskers and occurs only on the low frequency side of a critical band wide masker [4, 5]. In fact, the current opinion is that frequency components *remote* from the probe are primarily responsible for the effect, see e.g, [1].

There is also evidence that there is no difference between psychoacoustic tuning curves for 5 ms tones and narrow band noise sequences while the difference is clear between 35-ms signals [18]. This also supports the suggestion that the ear is quite insensitive to errors during the onset.

What makes it difficult to directly apply any above mentioned result to coding of a sudden transient is the diversity: it may have narrow or wide band, it may be an onset of a high level continuous sound or just a brief attack, e.g, of a percussive instrument, or even a start of a periodic pulse train, e.g., of a vowel in speech. Therefore, a set of different natural music or speech signals are used in the following listening tests.

3. A LOW-DELAY CODEC AND LISTENING TESTS

The coding scheme is based on frequency-warped linear predictive coding (WLP) techniques [16, 7, 10]. The decoder is a 20th order IIR-type warped lattice filter introduced in [6]. The continuously time-varying coefficients of the filter are estimated using a modified version of so called lattice method of linear prediction [13]. The technique is closely related to the gradient adaptive lattice method used in [10]. The encoder consists of an inverse filter, i.e., a warped FIR-type lattice filter which produces a residual signal which is transmitted to the decoder. A continuously varying filter was chosen in order to remove problems of frame-based processing. The coding and transmission of the filter coefficients is out of the scope of the current paper.

In [13], Makhoul introduced a set of techniques for estimating a reflection coefficient k_l of a lattice filter from its forward and backward residuals $f_l(n)$ and $b_l(n)$ at stage l. Using the *Harmonic Mean Method*, k_l for a signal sequence of L samples is given by

$$k_l = -\frac{2\sum_{k=0}^{L} f_{l-1}(k)b_{l-1}(k-1)}{\sum_{k=0}^{L} f_{l-1}^2(k) + b_{l-1}^2(k-1)} \quad (1)$$

In a block analysis technique by Makhoul and Viswanathan [14], this is used so that k_1 is computed first, then error signals $f_1(n)$ and $b_1(n)$, where $n = 0, 1, \cdots, L$, are produced using this value and k_2 is obtained using (1). This is repeated for all stages of the lattice filter.

In the current paper, a low delay coding scheme is simulated so that all the error signals

$$C(n) = f_{l-1}(k)b_{l-1}(k-1) \quad (2)$$

$$D(n) = f_{l-1}^2(k) + b_{l-1}^2(k-1) \quad (3)$$

of the lattice filter are first computed for the whole test signal and $l = 1$. Then, a sliding window function shown in Fig. 1 is applied to signals $C(n)$ and $D(n)$ to produce smoothed error sequences $\tilde{C}(n)$ and $\tilde{D}(n)$. After that a continuously time-varying path for the coefficient $k_1(n)$ is computed as

$$k_1(n) = -\frac{\tilde{C}(n)}{\tilde{D}(n)}. \quad (4)$$

The process continues so that the error sequences are computed for the second stage of the filter and the path for $k_2(n)$ is obtained. This is repeated for all the stages of the filter.

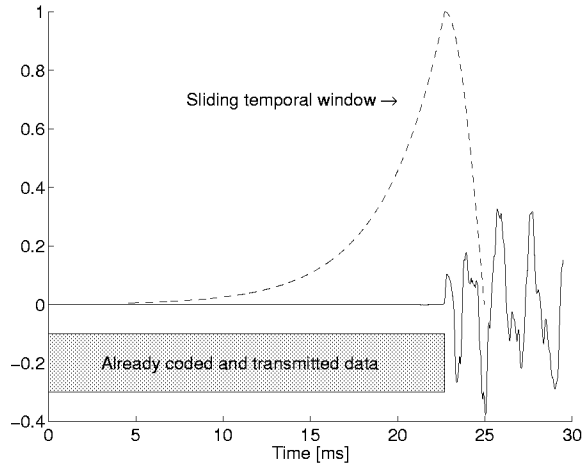

Figure 1: The position of the sliding temporal window function, which is used in estimating filter coefficients, and the coding process at onset of test sequence 4. The right end of the shaded bar shows the position where quantization occurs. The right part of the window function is a half of a cosine window and the left part is an exponentially decaying function.

Next, the time varying filter sequence was shifted so that the system mimics the performance of a codec having a coding delay of 50 sample periods (1.1 ms), i.e, the right end of smoothing window function proceeds 50 samples ahead of the time instant where the inverse filtering occurs and the quantized samples of the residual are transmitted to the decoder, see Fig. 1.

The performance of the spectral estimation technique is illustrated in Fig. 2. In each sub-figure, the lower curve is the waveform of the signal and the upper smoothed envelope shows the prediction gain as a function of time. In the very beginning of an onset the prediction gain is weak and, evidently, the estimated all-pole model of the signal is inaccurate because only a portion of the window function falls on the starting signal.

In the listening test the objective was to determine the detectability of coding errors after the onset in a brief interval of d milliseconds. The test sequences were produced so that white noise was added to the residual signal and the time-varying synthesis filter was driven using this as the excitation. After that, the interval d in the original signal was replaced by the deteriorated reconstructed signal, using cross-windowing, so that the signal energy within this interval was held constant. The subjects task was to use the method of adjustment to find the Signal-to-noise ratio (SNR) for the interval d of the residual signal so that the coding error is at the threshold of masking. Therefore, the detectability of coding error depends both on the accuracy of the all-pole model in the beginning of the signal and the amount and distribution of noise energy.

The test sequences are listed in Table 1. The sequences 2-6 are from the *McGill University Master Sample* CD-collection [15], 8 is the last hit of the *Castanets* test sequence commonly used in the MPEG community, and 9 is a synthetic harmonic signal having 10 sinusoids within the range of 80 to 4000 Hz. Sequences 1 and 7 are

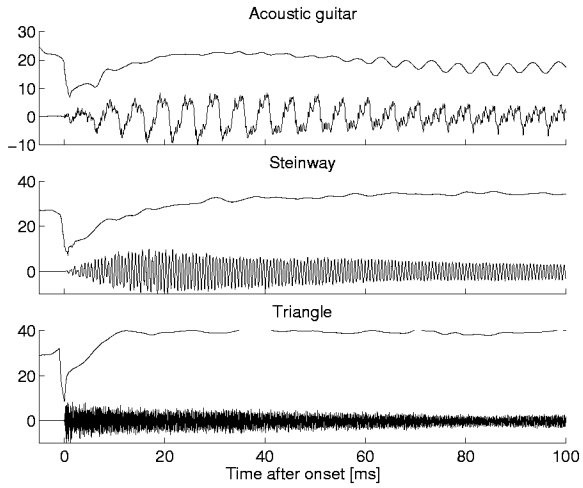

Figure 2: The waveforms of three test signals. The upper curve in each sub-figure shows the prediction gain [dB] as a function of time in the current low-delay coding experiment.

from collections of the Laboratory of Acoustics and Audio Signal Processing of Helsinki University of Technology. The duration of each test signal is 500 ms. All sequences except the *Castanets* were recorded in an anechoic chamber.

1	Acoustic guitar 3rd open	HUT/Acoustics
2	Marimba G2	McGill vol 3 04-03
3	Steinway piano F#6	McGill vol 3 01-70
4	Timbales	McGill vol 3 11-27
5	Triangle	McGill vol 3 12-25
6	Violin F4, pizzicato	McGill vol 1 03-12
7	Male speech, vowel	HUT/Acoustics
8	Castanets	a MPEG4 test sequence
9	Synthetic (10 harmonics)	

Table 1: Test sequences

4. RESULTS

Two different listening tests were performed using AKG K 240 headphones. Only one subject participated in the first experiment and five subjects participated in the second test. All the subjects are amateur musicians and are working in the Laboratory of Acoustics and Audio Signal Processing. Most of them are also experienced listeners. In the current paper, the term *threshold* is the SNR of the residual signal for which the coding error is just audible. That is, if the threshold is -20 dB, the excitation of the synthesis filter is practically random noise. Correspondingly, for the value of 0 dB, the energy ratio between original residual and noise is 1:1.

4.1. Experiment 1

In the first test the detectability of a corrupted signal segment, a *probe*, was measured as a function of the time position t_c of the probe relative to the onset of a signal. The duration of the probe was $d = 5$ ms. This experiment was necessary to eliminate the

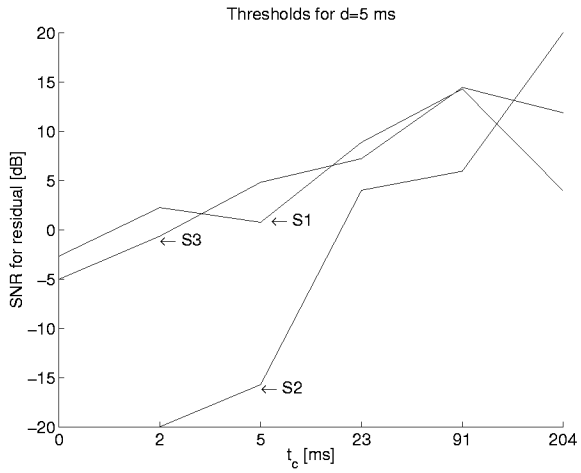

Figure 3: The SNR for residual at threshold for $d = 5$ ms and $t_c = 0, 2, 5, 23, 91, 204$ ms for signals 1, 2, and 3.

influence of the duration of the corrupted segment. Results for signals 1-3 support the hypothesis that even a remarkably corrupted onset may be accepted, see Fig 3. In particular, for the sequence 2, a completely random noise sequence may be used as excitation as long as the signal energy is held constant. For sequences 5-6 the trend is not that clear. For sequence 4 (Timbales) and 8 (Castanets) the threshold decreases as the position of the corrupted segment is farther away from the onset. This is probably due to that the signal has a very brief and intensive onset and rest of the signal is tonal and attenuates quickly, e.g, in sequence 8, the energy envelope decreases 20 dB during the first 5 ms. For the test sequence 7 it turned out that it is very difficult to determine the threshold precisely. However, for small values of t_c the threshold is below 0 dB. Sequence 9 showed a mild rising trend starting from the value of 5 dB.

4.2. Experiment 2

In the second experiment the detectability of a deteriorated onset as a function of the duration d was measured so that $t_c = 0$. Five subjects participated the test and the required SNR for the residual signal was determined for signal durations of $d = 2, 5, 9$, and 23 ms.

Mean data are presented in Fig. 4a. The thresholds increase monotonically for most of the test sequences. The average over all signals and all subjects for the values $d = 2, 5, 9$, and 23 ms are -12, -3, 4, and 8 dB, respectively.

5. DISCUSSIONS

One possible source of error in Experiment 2 is that the loudness of a noise burst or gap is a function of the duration of the event. A 2 ms noise burst presented in continuous noise is just perceptible if the difference between the level of the burst and the noise is approximately 7 dB. For a 23 ms burst this just noticeable difference is 4 dB [11]. Just perceptible level differences for 5 and 23 ms gaps differ only by 6 dB. This suggests that there may be a systematic error in the results of Experiment 2, but it is small (4-6 dB) compared to the difference observed in the listening test (10-20 dB).

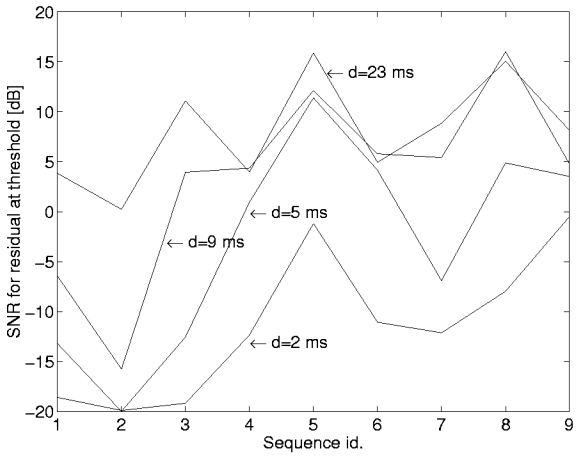

Figure 4: Mean data for the test sequences for four values of d.

In sequence 8, the threshold is low for a 2 ms probe but surprisingly high for a 5 ms and longer probes. The energy envelope of this signal decreases approximately 20 dB during the first 5 ms. In our experiments the first 5 ms are replaced with a steady noise segment with the same energy. Therefore, the error in the onset is relatively large in the temporal envelope of the signal. This indicates that it is not sufficient to reproduce only the energy and rough spectral envelope within this interval but also the temporal envelope of the onset.

For sequences 5 and 9, the threshold is at relatively high level for all the durations. Both signals have very clean and tonal onset with few distinct spectral peaks. In a pilot experiment it turned out that the ear is very sensitive to coding errors in the onset of a single tone. As mentioned in section 2 the overshoot effect may disappear in the case of single tone maskers. However, for both of the test signals the SNR for residual is below 0 dB for the shortest duration of the probe.

6. CONCLUSIONS

It seems plausible that the overshoot effect of hearing may be utilized in designing perceptual low-delay coding techniques. The threshold for all natural test sequences is very low during the first 2-5 ms. Hence, it is possible to reduce the bit-rate of the residual significantly during the very onset of a sound assuming that the spectrum and energy envelope of the reproduced signal approximately coincide with the original signal. In place of the residual, one may transmit more *forward* information which benefits the coding process after that interval, or transmit more accurate *spatial* information in the case of joint stereo coding [8].

However, due to the diversity of test sequences it is impossible to draw any explicit rule for how roughly spectral and temporal details can be reproduced during the first few milliseconds of the onset. Alternatively, we could use more synthetic test signals, e.g., noise and pure tone sequences, but it is already well documented how the masked threshold behaves in those cases and those signals are also quite uninteresting in any practical application.

The problem how to detect and choose onsets in continuous audio material requires more listening tests and experiments with a fully implemented low-delay audio codec.

7. ACKNOWLEDGMENT

This work has been supported by the Academy of Finland. The authors are grateful to the listening test subjects.

8. REFERENCES

[1] R P Carlyon and L J White. Effect of signal frequency and masker level on the frequency regions responsible for the overshoot effect. *J. Acoust. Soc. Am.*, 91(2):1034–1041, 1992.

[2] J-H Chen and R V Cox. The creation and evolution of 16 kbit/s ld-celp. *Speech Communication*, 12(2):103–111, 1993.

[3] H Fastl. Temporal masking effects: I. broad band noise masker. *Acustica*, 35(5):287–302, 1976.

[4] H Fastl. Temporal masking effects: II. critical band noise masker. *Acustica*, 36(5):317–330, 1977.

[5] H Fastl. Temporal masking effects: III. pure tone masker. *Acustica*, 43:282–294, 1979.

[6] A Härmä. Implementation of recursive filters having delay free loops. In *Proc of ICASSP'98*, volume III, pages 1261–1264, Seattle, 1998.

[7] A Härmä, U K Laine, and M Karjalainen. Warped linear prediction in audio coding. In *Proc. of IEEE Nordic Signal Processing Symposium NORSIG'96*, Helsinki, 1996.

[8] A Härmä, U K Laine, and M Karjalainen. An experimental audio codec based on warped linear prediction of complex valued signals. In *Proc. of ICASSP'97*, volume 1, pages 323–327, Munich, 1997.

[9] A Härmä, U K Laine, and M Karjalainen. WLPAC – a perceptual audio codec in a nutshell. In *AES 102nd Conv. preprint 4420*, Munich, 1997.

[10] A Härmä, U K Laine, and M Karjalainen. Backward adaptive warped lattice for wideband stereo coding. In *Proc. of EUSIPCO'98*, Rhodes, 1998.

[11] R J Irwin and S C Purdy. The minimum detectable duration of auditory signals for normal and hearing-impaired listeners. *J. Acoust. Soc. Am.*, 71(4):967–974, 1982.

[12] P Iverson and C L Krumhansi. Isolating the dynamic attributes of musical timbre. *J. Acoust. Soc. Am.*, 94(5):2595–2603, 1993.

[13] J Makhoul. Stable and efficient lattice methods for linear prediction. *IEEE Tr. ASSP*, pages 423–438, 1977.

[14] J Makhoul and R Viswanathan. Adaptive lattice methods for linear prediction. In *Proc. ICASSP'78*, pages 83–86, 1978.

[15] F Opolko and J Wapnick. *McGill University Marter Samples User's Manual*. McGill University Faculty of Music, Montreal, 1989.

[16] H W Strube. Linear prediction on a warped frequency scale. *J. of the Acoust. Soc. Am.*, 68(4):1071–1076, 1980.

[17] R von Klitzing and A Kohlrausch. Effect of masker level on overshoot in running- and frozen-noise maskers. *J. Acoust. Soc. Am.*, 95(4):2192–2201, 1994.

[18] D L Weber and R D Patterson. Sinusoidal and noise maskers in simultaneous and forward masking. *J. Acoust. Soc. Am.*, 75(3):925–931, 1984.