Audio Engineering Society
Convention Paper
Presented at the 111th Convention
2001 September 21–24 New York, NY, USA

Audio Coding Using Perceptually Controlled Bitstream Buffering

Christof Faller
Media Signal Processing Research, Agere Systems
600 Mountain Avenue, Murray Hill, New Jersey 07974, USA
cfaller@agere.com

ABSTRACT

Perceptual audio coders use a varying number of bits to encode subsequent frames according to the perceptual entropy of the audio signal. For transmission over a constant bitrate channel the bitstream must be buffered. The buffer must be large enough to absorb variations in the bitrate, otherwise the quality of the audio will be compromised. We present a new scheme for buffer control of perceptual audio coders. In contrast to conventional schemes the proposed scheme systematically reduces the variation in a perceptual distortion measure over time. The new scheme applied to a perceptual audio coder (PAC) improves the quality of the encoded signal for a given buffer size. The same technique can be used to increase the performance of other coders such as MPEG-1 Layer III or MPEG-2 AAC while maintaining backward compatibility.

1 INTRODUCTION

Typical non-stationary signals such as audio signals have a variable inherent perceptual entropy [1]. To approach the compression limit (i.e. perceptual entropy for transparent audio coding) of such signals, variable bitrate compression techniques are used. Perceptual audio coders [2, 3] quantize the spectral components of an audio signal such that the quantization noise follows the noise threshold determined by the perceptual model. For transparent audio coding the perceptual model computes the masked threshold [4]. For non-transparent audio coding the parameters of the perceptual model can be tuned such that it generates a supra-threshold, i.e. a noise threshold which is above the masked threshold. In this case, more quantization noise is introduced into the encoded audio signal and the bitrate will be lower.

In a general sense, a perceptual model is an algorithm which computes for each frame k of an audio signal a noise threshold for each spectral component of this signal with the perceived distortion $D[k]$ as a model parameter. Ideally the computed noise threshold has the property that it is the spectrally shaped noise with the largest possible variance for a perceived distortion of $D[k]$. For $D[k] = 0$ the perceptual model computes the masked threshold (for transparent audio coding) and for $D[k] > 0$ it computes a supra-threshold (for non transparent audio coding). Ideally a perceptual model computes the noise thresholds such that for a constant parameter $D[k]$ the perceived distortion of the entire coded audio signal is constant.

In this paper, we assume that an audio coder without rate control or buffer control is the ideal case of encoding an audio signal. In this case the audio signal is encoded with quantization noise equal to the noise threshold given by the perceptual model with the distortion held constant $D[k] = D_R$. We call the frame bitrates $M[k]|_{D_R}$ in this case *inherent bitrates* with an average bitrate of R. The average bitrate

$$R = \frac{1}{N} \sum_{k=1}^{N} M[k]|_{D_R} \tag{1}$$

of an audio signal encoded with a constant distortion $D[k] = D_R$ is not known prior to encoding the signal. The number of frames to be encoded is N.

For a specific desired bitrate R the audio signal is ideally encoded with frame bitrates $M[k]$ $(bits/frame)$ equal to the inherent bitrates $M[k]|_{D_R}$. A criterion for the optimality of the encoding process is

$$\sigma_D^2 = E\{(D[k]|_{M[k]} - D_R)^2\}\,. \tag{2}$$

The smaller σ_D^2 is the less is the distortion varying over time with no variation at all for $\sigma_D^2 = 0$. The more the distortion $D[k]$ differs from D_R the more the bitrate $M[k]$ differs from the inherent bitrate $M[k]|_{D_R}$ and a related measure for optimality is

$$\sigma_R^2 = E\{(M[k] - M[k]|_{D_R})^2\}\,. \tag{3}$$

In the optimal case when an audio signal is encoded with a constant distortion $D[k] = D_R$ the bitrate of each frame $M[k] = M[k]|_{D_R}$ will vary significantly as shown in Fig. 1. However, many applications require a constant bitrate R transmission from the encoder to the decoder. To operate an audio encoder at a constant bitrate one can iteratively encode each frame k with various distortions $D[k]$ until the bitrate of the frame $M[k]$ is equal to the desired bitrate R as shown in Fig. 2. In this case the bitrates $M[k] = R$ differ significantly from the inherent

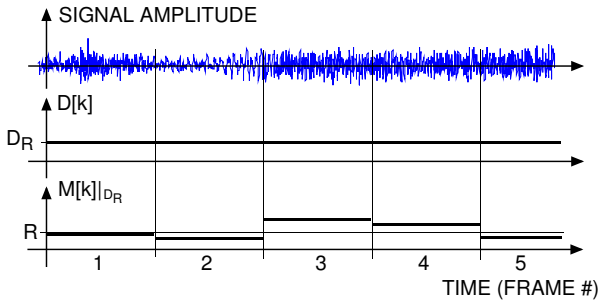

Fig. 1: Perceptual audio coders are by nature variable bitrate source coders. For encoding an audio signal (top) at a constant distortion $D[k]$ (middle) each frame has a different bitrate $M[k]$ (bottom).

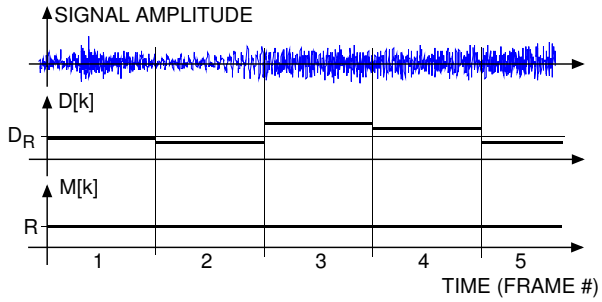

Fig. 2: For encoding an audio signal (top) at a constant bitrate $M[k] = const.$ (bottom) the distortion $D[k]$ (middle) must be varied in time.

bitrates $M[k]|_{D_R}$ leading to a large σ_R^2 and to significant variations of the distortion $D[k]$ over time. The encoded audio signal has now a significantly worse quality than in the ideal case of encoding with $M[k] = M[k]|_{D_R}$ for the same average bitrate R.

Figure 3 shows conceptionally these two extreme cases of operation for an average bitrate of R. In one case (a) only the bitrate is varied ($M[k] = M[k]|_{D_R}, D[k] = const.$) and in the other case (b) only the distortion is varied ($M[k] = R, D[k]$).

Many applications for constant bitrate transmission, such as digital radio broadcasting [5, 6] or Internet streaming [7], can afford an end-to-end delay. For such applications the bitstream of the audio coder can be buffered. The buffer can absorb a certain degree of variation in the bitrate $M[k]$, leading to reduced variations in the distortion $D[k]$ over the case of an audio coder without bitstream buffering. A tradeoff can be made between variation in bitrate and variation in the distortion. The smaller the variation in the distortion, the larger the variation in the bitrate, and visa versa. The buffer size and

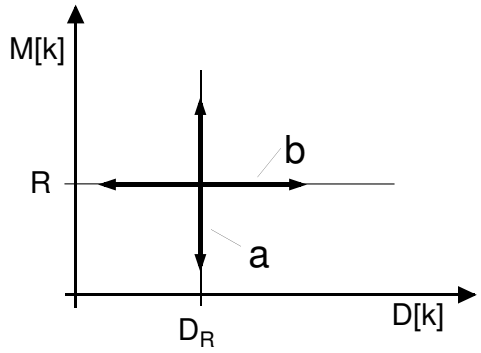

Fig. 3: The two extreme cases of operation of an audio coder: (a) ideally the audio coder is unconstrained and only varying the bitrate $M[k]$, (b) the audio coder forced to be a constant bitrate coder by varying only the distortion.

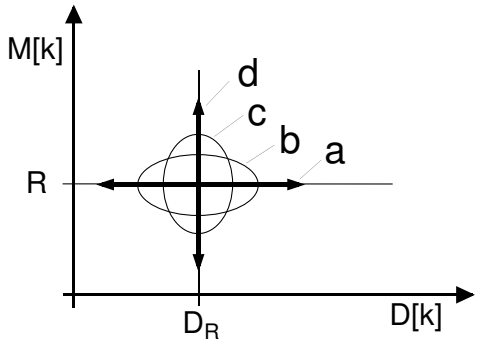

Fig. 4: Different scenarios for audio coding over a constant bitrate transmission channel: (a) constant bitrate audio coding without bitstream buffering, (b) buffered bitstream, (c) buffered bitstream with larger buffer or better buffer control, (d) infinitely large buffer.

the buffer control scheme determine how much variation in the bitrate can be absorbed by the buffer. Therefore the buffer size and the buffer control scheme also determine the amount of variation in the distortion. Figure 4 schematically shows the regions in which the bitrates $M[k]$ and distortions $D[k]$ lie for different scenarios:

a) Constant bitrate audio coding $M[k] = R$ (no bitstream buffering).

b) Audio coding with bitstream buffering.

c) Audio coding with bitstream buffering (larger buffer or better buffer control scheme than b).

d) Unconstrained audio coding $M[k] = M[k]|_{D_R}$ (infinitely large buffer).

Figure 5 shows an audio encoder and decoder with a buffered bitstream for a constant bitrate transmission of the bits. In this scenario, the $M[k]$ bits of the encoded frame, at the time of each frame k, are put into a FIFO (first-in-first-out) buffer while R_d bits are removed from the FIFO buffer by the constant bitrate transmission channel. The number of data bits in the encoder buffer can be expressed iteratively as

$$l[k] = l[k-1] + M[k] - R_d, \tag{4}$$

with an initial buffer level of $l[0] = l_d$ *bits*. A *buffer control* scheme monitors the buffer level $l[k]$ and influences the encoding process to make sure the buffer does not overflow. Buffer underflow can be easily prevented by padding additional (non-used) bits to the frame when underflow would occur.

For many applications the size of the buffer is restricted by a desire for a small end-to-end delay or by cost and

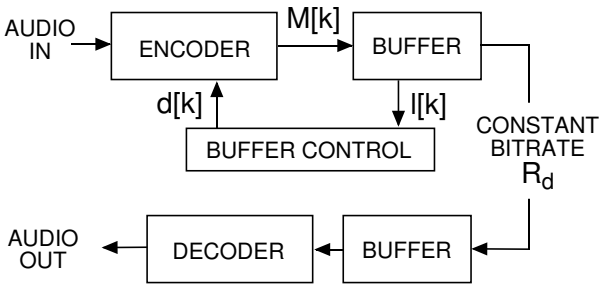

Fig. 5: An audio encoder and decoder with a constant bitrate transmission channel.

complexity restrictions of the encoder or decoder. For broadcasting applications the desired end-to-end delay is limited by the cost of the decoder and the tune-in time, i.e. the time it takes from a request for playback until the audio actually plays back. Therefore there is a need of a buffer control scheme which minimizes the variation in the distortion for a given limited buffer size. In this paper, we present a novel buffer control scheme for audio coding which significantly reduces the variation in the distortion $D[k]$ for a given buffer size.

In Section 2, we review conventional buffer control schemes such as used in MPEG-1 Layer III [8], MPEG-2 AAC [3], or PAC [2]. In Section 3, we present the new scheme for buffer control with significantly reduced variation in the distortion over time. The results of a subjective listening test are summarized in Section 4. Some conclusions are drawn in Section 5.

2 CONVENTIONAL BUFFER CONTROL

Typically buffer control schemes for audio coders encode

each frame k with two processing loops [3, 8, 2, 9]. The *outer loop* determines for each frame k a momentary bitrate $M_d[k]$ at which it should be encoded. The momentary bitrate $M_d[k]$ is computed as a function of the buffer level $l[k-1]$ and the perceptual entropy $M_0[k]$ [1] of the frame or a related measure. The *inner loop* iteratively re-encodes the frame at different levels of distortion until the bitrate of the encoded frame $M[k]$ is close to $M_d[k]$.

Typically the outer loop determines the bitrate of each frame $M_d[k]$ with a strategy of keeping a fairly low buffer level (low buffer level = many bits available) in order to have plenty of bits available in the case of frames with very high inherent bitrate. For example frames with transients usually have a very high inherent bitrate compared to frames without transients.

A simpler strategy would be to assign to each frame a bitrate of R_d, i.e. operating the audio coder at a constant bitrate. But by having an outer loop assigning to each frame an individual bitrate $M_d[k]$ the audio coder can take advantage of the variability in bitrate which is possible because of the buffering of the bitstream.

The outer loop determines heuristically the bitrate $M_d[k]$ for each frame without considering the effect it has on the distortion. The distortion is determined independently by the inner loop. In contrast to our proposed scheme iterative schemes such as described in this section do not aim at explicitly reducing the local variation in the distortion $D[k]$ and are in that sense far from optimal.

3 PERCEPTUAL BUFFER CONTROL

3.1 Optimal Audio Coding with an Average Desired Bitrate

In the optimal case of encoding an audio signal with a constant distortion $D[k] = D_R$ the average bitrate R (1) is unknown prior to encoding the whole audio signal. For an average bitrate equal to a desired bitrate of R_d one can encode the audio signal iteratively for different distortions until the average rate R (1) is equal to the desired bitrate R_d. Figure 6 shows schematically the average bitrate R (1) as a function of the constant distortion $D[k] = D_R$ and the point (D_{R_d}, R_d) at which the signal is encoded.

The method described is suitable for encoding audio signals in cases when the whole signal is given at once and if there is no buffer constraint. Applications are storage of audio signals. However the method described is not suitable for applications where the entire signal is not available before encoding (e.g. in real-time applications or applications with limited signal buffers).

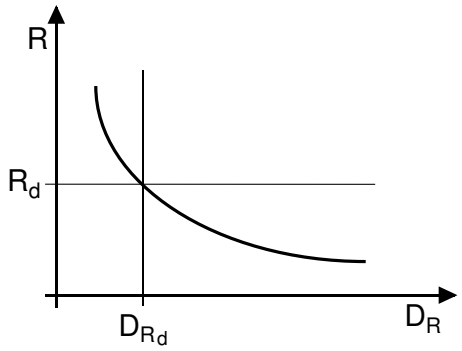

Fig. 6: The average bitrate R as a function of the constant distortion $D[k] = D_R$.

3.2 Real-Time Audio Coding with an Average Desired Bitrate

The goal is to approximate the ideal case of encoding the audio signal with a constant distortion D_{R_d}. Without introducing any additional delay in the audio coder, at the time of frame k only frames $k, k-1, k-2, \ldots$ are given. Instead of considering the average bitrate over the whole audio signal (1) the average bitrate is estimated locally in time,

$$R[k] = \sum_{i=k-W+1}^{k} w[i] M[i]|_{D_R[k]}\,, \qquad (5)$$

where $w[i]$ is the estimation window having a time span of W frames.

Each frame k of the audio signal is encoded with a distortion $D_{R_d}[k]$ such that the estimated average bitrate $R[k]$ (5) is equal to the desired bitrate R_d. For each frame k the distortion $D_{R_d}[k]$ can be computed iteratively by encoding the audio signal within the window $w[i]$ for different distortions until the estimated average rate $R[k]$ (5) is equal to the desired bitrate R_d.

The method described is suitable for real-time applications since it does not require any lookahead.

3.3 Real-Time Audio Coding with a Buffer Constraint

If for each frame the distortion is chosen to be $D_{R_d}[k]$ as described in the previous section then the expected long-term average bitrate of the audio coder is R_d. However the variance of the buffer-level is monotonically increasing over time. If we assume that

$$e_M[k] = M[k]|_{D_{R_d}[k]} - R_d \qquad (6)$$

is an independent and identically distributed (i.i.d.) random variable with a variance of σ^2, then the buffer level

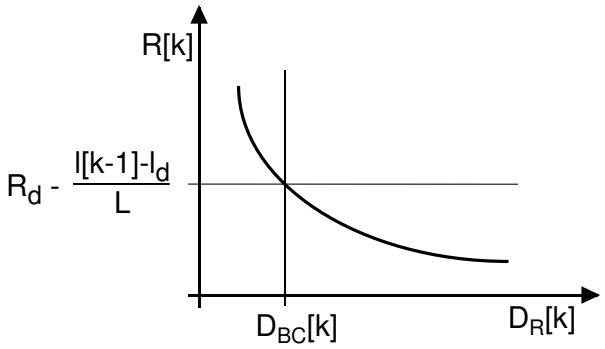

Fig. 7: The average bitrate $R[k]$ as a function of the distortion $D_R[k]$ and the point at which a frame is encoded.

(4) is the sum of k i.i.d. random variables with a total variance of $k\sigma^2$.

To encode the audio signal such that the variance of the buffer level has an upper bound the distortion for each frame $D_{BC}[k]$ is chosen such that the estimated average bitrate $R[k]$ (5) is equal to

$$R_{BC}[k] = R_d - \frac{l[k-1] - l_d}{L}, \tag{7}$$

where L determines the weighting of the buffer level deviation on the chosen average bitrate in (5). Each frame has an expected bitrate of $R_{BC}[k]$ instead of the desired bitrate R_d. Thus the buffer-level is statistically driven to the desired buffer-level l_d with a time constant of LT *seconds*. T is the duration of one frame in seconds. For our experiments we chose $L = 50$. Figure 7 shows the estimated average bitrate $R[k]$ (5) as a function of the distortion $D[k] = D_R[k]$ and the point at which a frame is encoded.

We now show that when the audio signal is encoded with distortions $D_{BC}[k]$, the mean of the buffer-level $E\{l[k]\}$ is l_d and the variance $\sigma^2_{l[k]}$ is upper bounded by

$$\sigma_e^2 \frac{1}{1-(1-\frac{1}{L})^2}\,, \tag{8}$$

where σ_e^2 is $E\{e^2[k]\}$ with

$$e[k] = M[k] - (R_d - \frac{l[k-1] - l_d}{L}). \tag{9}$$

The variable $e[k]$ is assumed to be i.i.d. with zero mean. For the derivation of the mean $E\{l[k]\}$ and the bound for the variance (8) we re-write the buffer-level (4) with (9) as

$$l[k] = 1 - \frac{1}{L} l[k-1] + \frac{1}{L} l_d + e[k]\,. \tag{10}$$

With an initial buffer-level of $l[0] = l_d$ and the first frame to be encoded $k = 1$, (10) written non-iteratively is

$$l[k] = l_d + \sum_{i=1}^{k} e[i](1-\frac{1}{L})^{k-i}\,. \tag{11}$$

Equation (11) and considering that $e[k]$ has zero mean yields

$$E\{l[k]\} = l_d\,, \tag{12}$$

and the variance $\sigma^2_{l[k]}$ as a function of k is

$$\sigma^2_{l[k]} = E\{(l[k]-l_d)^2\} = \sum_{i=1}^{k} \sigma_e^2 (1-\frac{1}{L})^{2(k-i)}\,. \tag{13}$$

Given (13) one can easily show that the variance of the buffer-level converges to the value given in (8).

3.4 Efficient Implementation

In this section, we describe a scheme for efficient implementation of the buffer control scheme described in Section 3.3. Similarly the rate control schemes described in Sections 3.1 and 3.2 can be implemented efficiently.

The buffer control scheme (Section 3.3) needs to find for each frame k the solution of (5) for $R[k] = R_{BC}[k]$ (7). For each frame k we approximate the function f_k which maps the distortion $D_R[k]$ to the estimated average bitrate $R[k]$ (5) (figure 7),

$$R[k] = f_k(D_R[k])\,, \tag{14}$$

by linearly interpolating between a set of computed discrete points. The discrete points are obtained by computing the estimated bitrates $\{R_i[k]\}$ given a set of predefined distortions $\{D_i\}$ (5),

$$R_i[k] = \sum_{i=k-W+1}^{k} w[i] M[i]|_{D_i}\,, \tag{15}$$

with $i \in \{1, 2, \ldots, I\}$. Figure 8 shows an example of the approximation of f_k given the discrete points (R_i, D_i). Given f_k frame k is encoded with a distortion of

$$D_{BC}[k] = f_k^{-1}(R_{BC}[k])\,. \tag{16}$$

Each frame k of the audio signal is encoded with this algorithm:

1. Encode frame k for each of the I distortions D_i to compute the frame bitrate $M[k]|_{D_i}$.
2. Estimate the average bitrate $R_i[k]$ for each distortion D_i (15) given current and past frame bitrates.
3. Interpolate between the values $(R_i[k], D_i)$ to obtain an approximation of the function f_k (Fig. 8).

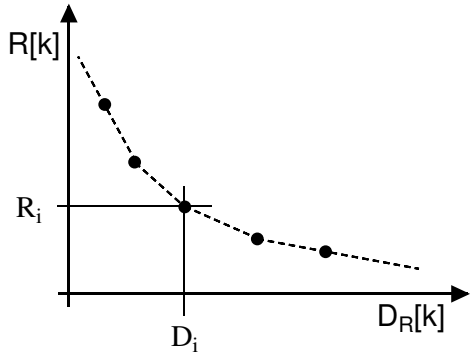

Fig. 8: The average bitrate $R[k]$ as a function of the distortion $D_R[k]$ is approximated with linear interpolation.

4. Encode the frame with a distortion of $D_{BC}[k]$ (16).

The number of coding iterations for each frame is $I+1$.

We accurately computed the estimated average bitrate $R[k]$ (5) as a function of the distortion $D_R[k]$ for PAC for a wide variety of audio signals. Figure 9 shows that for PAC [2] the function f_k can be accurately approximated with a straight line with a time-invariant slope q within the range of operation used. Therefore f_k can be approximated by just computing one point $(D_1, R_1[k])$,

$$R[k] = f_k(D_R[k]) \approx q(D_R[k] - D_1) + R_1[k]\,. \qquad (17)$$

The number of coding iterations for encoding each frame of PAC is only 2 ($I = 1$). Therefore the new scheme is significantly less complex than PAC's previous iterative scheme. PAC's previous iterative scheme requires significantly more coding iterations for each frame to be encoded.

4 EXPERIMENTAL RESULTS

We compared the performance of PAC [2] using its conventional iterative scheme (Section 2) and the new buffer control scheme (Section 3) for a wide variety of speech and music clips. For that purpose we encoded a set of 57 stereo clips with a total length of 48 min. The samplerate of the clips was 32 kHz. The buffer size was chosen for an additional delay of 180 ms.

Figure 10 shows the histogram of the inherent bitrates of each of the 57 clips for one common distortion $D[k] = D_{R_d}$. The average inherent bitrates of the individual clips differ by more than $\pm 20\%$ from R_d. For applications such as digital radio broadcasting or Internet streaming the audio coder has to encode audio material with greatly varying inherent bitrates (relative to a constant distortion $D[k]$). For testing the robustness of

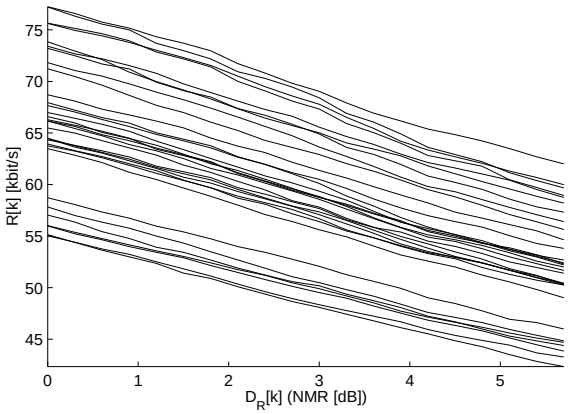

Fig. 9: The estimated average bitrate $R[k]$ as a function of the distortion $D_R[k]$ for PAC.

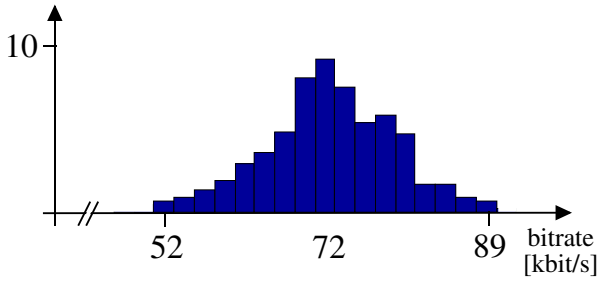

Fig. 10: Bitrates corresponding to the inherent bitrates of the 57 audio clips for one common distortion $D[k] = D_{R_d}$.

the scheme against these variations we chose a group of 3 clips out of the 57 clips for each of the following three classes:

Class A: Average inherent bitrate is close to the desired bitrate.

Class B: Average inherent bitrate is significantly larger than desired bitrate.

Class C: Average inherent bitrate is significantly smaller than desired bitrate.

Figure 11 shows the distortion as a function of time for PAC's previous iterative scheme and the new scheme for a signal of each of the three classes. As expected the new scheme has significantly less variation in the distortion (noise-to-masked ratio, NMR) over time than the PAC's previous iterative scheme. Also the new scheme is more robust in the sense that its behavior is very similar for each of the three classes.

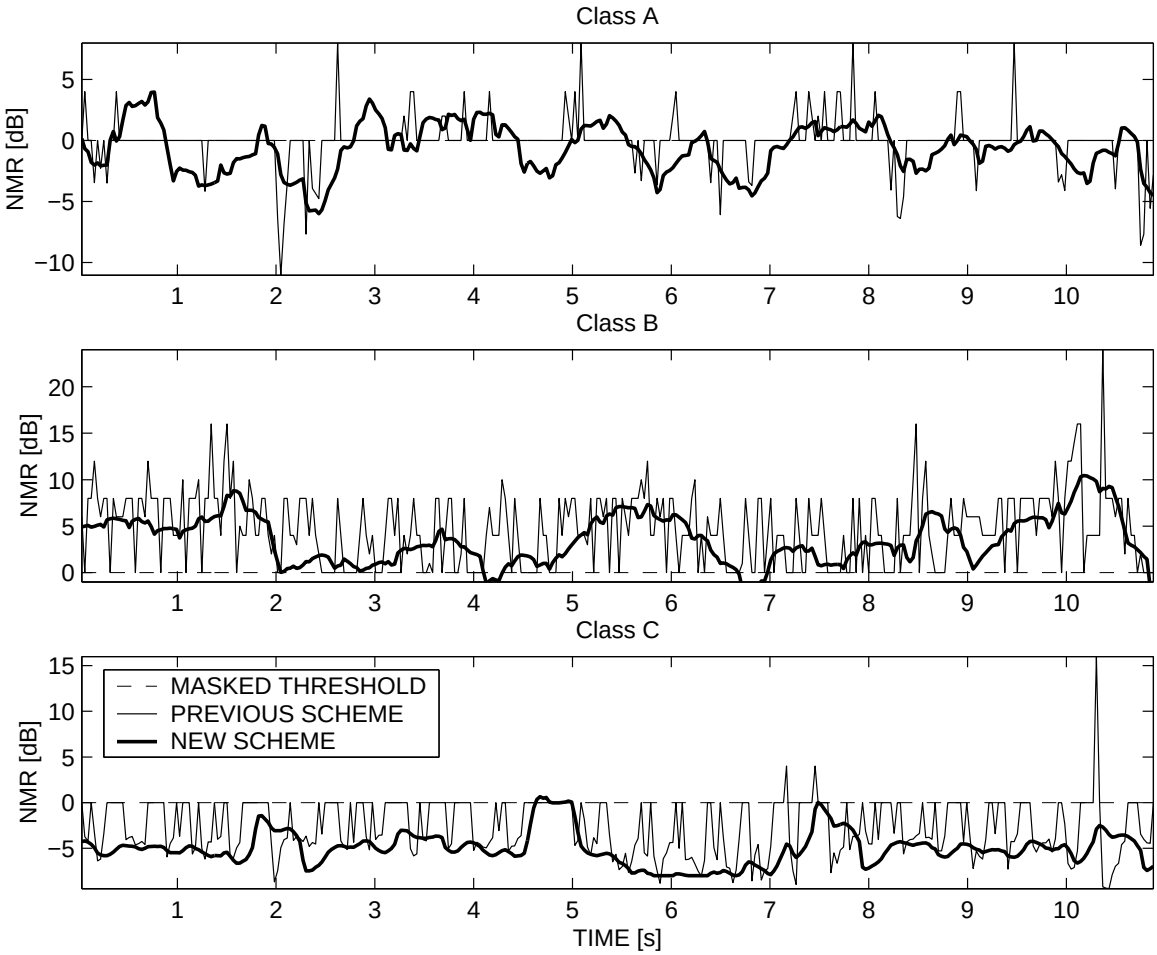

Fig. 11: Examples of the distortion of the three classes of clips (A, B, C) relative to the masked threshold for the new scheme and PAC's previous iterative scheme.

We conducted a blind triple stimulus test with a seven-grade comparison scale [10]. The test was carried out in a sound booth with the signals presented to the listeners with Stax headphones. The eight listeners were presented with a triple of signals, each of 10 s length for each trial. The uncoded source signal (reference) was presented first followed by the coded clips of the previous iterative and new scheme in random order. The quality difference of the coded items was graded with respect to the reference using the seven-grade comparison scale [10].

The average subjective comparison scores of both schemes for the three clips of each class are shown in Fig. 12. The quality of the coded clips improved significantly for class B. Class B are the clips that are most difficult to encode because their inherent bitrate is much higher than the desired bitrate.

5 CONCLUSIONS

In this paper, we presented a new paradigm for buffer control for audio coding. Instead of having an inner and outer loop heuristically determining the distortion for each frame, the new scheme systematically reduces the local variation in the distortion by encoding each frame with a distortion based on statistical bitrate estimations.

The new scheme's behavior is not dependent on how much the average inherent bitrate diverges from the desired bitrate. Therefore with the new scheme the encoder performs equally well for a wide range of signals. Also the new scheme is suitable for making audio-coders more

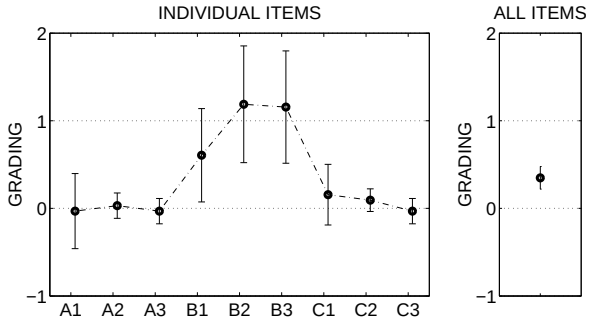

Fig. 12: Relative grading of the new scheme versus PAC's previous iterative scheme. A seven-grade comparison scale was used [10].

scalable in the sense that they operate well over a larger range of bitrates.

The proposed scheme is significantly less complex than iterative schemes. For each frame only two coding iterations need to be carried out. For previous iterative schemes the average number of coding iterations for a frame is significantly higher. Additionally the computational complexity of the new scheme is varying much less in time because the number of coding iterations is the same for each frame. Because of the lower complexity and less variation of complexity in time the new scheme can be implemented in real-time with a more modest processor and a smaller jitter buffer than previous iterative schemes.

It is shown in this paper that a statistical approach to bit-allocation for audio coding is possible, and superior to conventional iterative approaches. A subjective blind test has shown that the proposed scheme significantly improves perceptual audio coders with buffered bitstreams for a given buffer size.

The author thanks Frank Baumgarte, Tomas Gaensler, Peter Kroon, Sean Ramprashad, Gerald Schuller, and Martin Vetterli for valuable discussions and suggestions.

6 REFERENCES

[1] J. D. Johnston, "Estimation of perceptual entropy using noise masking criteria," in *ICASSP-88 Conference Record*, 1988.

[2] 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., chapter 42. CRC Press, IEEE Press, Boca Raton, Florida, 1997.

[3] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, and Y. Oikawa, "ISO/IEC MPEG-2 advanced audio coding," *J. Audio Eng. Soc.*, vol. 45, no. 10, pp. 789–814, 1997.

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

[5] R. K. Jurgen, "Broadcasting with digital audio," *IEEE Spectrum*, pp. 52–59, Mar. 1996.

[6] W. Pritchard and M. Ogata, "Satellite Direct Broadcast," in *Proc. IEEE*, July 1990, vol. 78, pp. 1116–1140.

[7] M. Dietz and K. H. Brandenburg, "Audio compression for network transmission," *J. Audio Eng. Soc.*, pp. 58–70, Jan./Feb. 1996.

[8] K. H. Brandenburg and G. Stoll, "ISO-MPEG-1 audio: a generic standard for coding of high-quality digital audio," *J. Audio Eng. Soc.*, pp. 780–792, October 1994.

[9] F.-R. Jean, C.-F. Long, T.-H. Hwang, and H.-C. Wang, "Two-stage bit allocation algorithm for stereo audio coder," in *IEEE Proc.-Vis. Image Signal Process.*, Oct. 1996, vol. 143, pp. 331–336.

[10] International Telecommunication Union ITU, *Rec. ITU-R BS.562.3*, 1990, http://www.itu.org.