AUDIO CODING WITH POWER SPECTRAL DENSITY PRESERVING QUANTIZATION

Minyue Li,[1] *Janusz Klejsa,*[1] *Alexey Ozerov,*[2] *and W. Bastiaan Kleijn*[1,3]

[1] School of Electrical Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

[2] Technicolor, 1 avenue de la Belle Fontaine, F-35576 Cesson Sevigne, France

[3] School of Engineering and Computer Science, Victoria University of Wellington, New Zealand

ABSTRACT

The coding of audio-visual signals is generally based on different paradigms for high and low rates. At high rates the signal is approximated directly and at low rates only signal features are transmitted. The recently introduced distribution preserving quantization (DPQ) paradigm provides a seamless transition between these two regimes. In this paper we present a simplified scheme that preserves the power spectral density (PSD) rather than the probability distribution. In a practical system the PSD must be estimated. We show that both forward adaptive and backward adaptive PSD estimation are possible. Our experimental results confirm that preservation of PSD at finite precision leads to a unified coding paradigm that provides effective coding at both high and low rates. An audio coding application shows the perceptual benefits of PSD preserving quantization.

Index Terms— Quantization, audio coding, predictive coding, forward adaptation, backward adaptation.

1. INTRODUCTION

To achieve good perceived quality for reproducing audio-visual signals, a natural requirement is to preserve features of the source signal. However, feature preserving coding procedures generally do not converge to transparent quality with increasing rate. On the other hand, coding approaches that lead to transparent quality at high rates are inefficient in preserving the features of the source signal when rate decreases. This has led to the usage of different coding paradigms for high and low coding rates. A recently developed paradigm that unifies high and low rate coding is distribution preserving quantization (DPQ) [1]. DPQ preserves the probability distribution of the source, thus retaining all features related to statistics of the signal. Subject to the distribution preserving constraint, DPQ minimizes a conventional distortion criterion. The result is a seamless transition from signal synthesis to signal quantization with increasing bit rate.

A mean-squared-error (MSE) optimized DPQ can be constructed by using a dithered quantizer and a non-linear transformation [2]. Unfortunately, the non-linear transformation, which is used to retrieve the source probability distribution, is computationally complex. In many cases, preserving the probability distribution exactly may not be feasible due to the associated complexity. This issue may be addressed by relaxing the distribution preserving constraint and maintaining only those statistical features that are perceptually most important. For example, in audio applications, it may be sufficient to preserve the power spectral density (PSD) of the source. This leads to a technique called PSD preserving quantization (PSD-PQ). PSD-PQ can be implemented by means of a pre/post-filtered dithered quantizer [1].

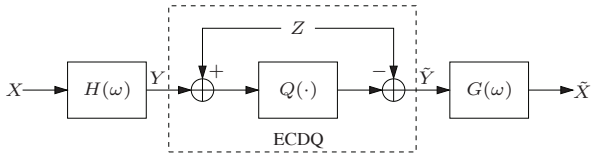

Fig. 1. Diagram of PSD-PQ.

In this paper we take a broader perspective than [1] and consider how the source PSD is conveyed to the decoder and the effect of the associated approximations. Two procedures are natural. First, the signal model can be extracted from the data, encoded, and transmitted to the decoder. This means that we have to consider the effect of model mismatch. An alternative solution is to extract the model from the most recent available decoded data at both the encoder and the decoder. In this paper we show that such a backward adaptive PSD-PQ converges to the optimal configuration over time.

To show the effectiveness of PSD-PQ we consider its application to audio coding. A listening test shows that PSD-PQ leads to better perceived quality than a comparable rate-MSE optimized quantizer.

2. THE PSD-PQ APPROACH

The PSD-PQ scheme considered in this paper follows [1]. Here, we extend our earlier results and consider the effect of model mismatch. In general, the PSD-PQ method is motivated from the fact that the source PSD can be retrieved by filtering the output of a dithered quantizer, which adds uncorrelated noise to the source signal. To enhance the rate-MSE performance, a pre-filter is also introduced as in [3].

2.1. System Structure

Let the source X be a stationary process with PSD $S(\omega)$ (we assume $S(\omega) > 0, \forall\omega$). We propose a coding system, shown in Fig. 1, consisting of a pre-filter $H(\omega)$, an entropy coded dithered quantizer (ECDQ) with subtractive dither Z and lattice quantizer $Q(\cdot)$, together with a post-filter $G(\omega)$. According to [4], ECDQ is effectively an additive noise channel, and the optimal achievable bit rate of the ECDQ, which is the entropy rate of its output given the dither, equals the mutual information rate between the input and the output of this channel. In addition, it is always possible to transform a lattice into a *white* lattice, for which the effective channel noise is white in the *wide sense* [4]. Here we also scale the lattice so that the power of the effective channel noise equals unity, leaving the rate-distortion tradeoff fully determined by the choice of the pre/post-filter.

For this coding scheme, the PSD of the reconstructed signal $\tilde{X}$ is $\tilde{S}(\omega) = |H(\omega)|^2|G(\omega)|^2 S(\omega) + |G(\omega)|^2$. Then a sufficient and

necessary condition for a system to be a PSD-PQ is

$$|H(\omega)|^2|G(\omega)|^2S(\omega)+|G(\omega)|^2=S(\omega). \quad (1)$$

2.2. Optimal Pre/Post-filter

In [1], the lowest possible rate for PSD-PQ of stationary Gaussian processes is derived for any MSE distortion level using an information theoretical argument. This optimal rate-MSE relation is given by the following proposition.

Proposition 1. *For any stationary Gaussian process, no PSD-PQ schemes can achieve a rate smaller than*

$$R=\frac{1}{4\pi}\int_0^{2\pi}\log\frac{(\lambda^2+4S^2(\omega))^{\frac{1}{2}}+\lambda}{2S(\omega)}d\omega, \quad (2)$$

if the MSE is smaller than

$$D=\frac{1}{2\pi}\int_0^{2\pi}2S(\omega)+\lambda-\left(4S^2(\omega)+\lambda^2\right)^{\frac{1}{2}}d\omega. \quad (3)$$

We will show that, using a lattice quantizer of infinite dimensionality and a proper pre/post-filter, the PSD-PQ in Fig. 1 can achieve this lower bound on the rate. It is known that, when applied to a stationary Gaussian process, the rate of an ECDQ with a pre-filter and a normalized infinite dimensional lattice is [3]

$$R^\star=\frac{1}{4\pi}\int_0^{2\pi}\log(|H(\omega)|^2S(\omega)+1)d\omega. \quad (4)$$

With a lattice of finite dimensionality, the bit rate of a pre-filtered ECDQ can be upper bounded by an offset up to $0.5\log(2\pi\mathrm{e}/12)\approx 0.254$ bits/sample from (4). Next, we provide the optimal pre/post-filter by the following proposition.

Proposition 2. *Suppose that the rate of a pre/post-filtered ECDQ achieves (4). Among all the pre/post-filters that preserve the source PSD, those that minimize the rate satisfy*

$$|H(\omega)|^2=\frac{(\lambda^2+4S^2(\omega))^{\frac{1}{2}}-\lambda}{2\lambda S(\omega)}, \quad (5)$$

and

$$G(\omega)=\lambda H^*(\omega), \quad (6)$$

where λ is chosen to achieve a certain MSE.

The design of the pre/post-filter facilitates a tradeoff between the rate and the MSE by adjusting λ. A derivation of this optimal pre/post-filter, which is based on variational calculus, is given in [1]. We can verify the optimality of this pre/post-filter. First, it is easy to see that the proposed pre/post-filter fulfills the PSD preserving condition (1). Then, substituting (5) into (4), we can also see that the rate coincides with (2), while the MSE [3]

$$D^\star=\frac{1}{2\pi}\int_0^{2\pi}|H(\omega)G(\omega)-1|^2S(\omega)+|G(\omega)|^2d\omega \quad (7)$$

equals (3).

3. SPECIFICATION OF THE PSD

PSD preservation requires prior knowledge of the source PSD, which is generally unavailable in practical applications. The PSD must be estimated from the signal and the estimated PSD must be made available to both encoder and decoder. To this end, either *forward* or *backward* adaptation can be used. We first discuss a generic property, which shows that the coder structure reduces errors in the PSD, then discuss forward adaption and backward adaptation, respectively.

3.1. PSD Mismatch

PSD estimation must be used and therefore, the PSD may not be exactly preserved. We now prove a nice property of the proposed PSD-PQ: the PSD of the reconstruction is always closer than the PSD estimate to the source PSD.

Proposition 3. *Let $\hat{S}(\omega)$ be an estimate of the source PSD $S(\omega)$, and $\tilde{S}(\omega)$ be the PSD of the reconstruction of the proposed PSD-PQ, for which the pre/post-filter are based on $\hat{S}(\omega)$. For any ω, one of the following conditions must be fulfilled:*

$$\begin{aligned}&0<\tilde{S}(\omega)-S(\omega)<\hat{S}(\omega)-S(\omega),\\ &\tilde{S}(\omega)-S(\omega)=\hat{S}(\omega)-S(\omega)=0,\\ &\hat{S}(\omega)-S(\omega)<\tilde{S}(\omega)-S(\omega)<0.\end{aligned} \quad (8)$$

Proof. With elementary algebra, we can obtain

$$\begin{aligned}\tilde{S}(\omega)&=|H(\omega)|^2|G(\omega)|^2S(\omega)+|G(\omega)|^2\\ &=S(\omega)+\frac{(\lambda^2+4\hat{S}^2(\omega))^{\frac{1}{2}}\lambda-\lambda^2}{2\hat{S}^2(\omega)}(\hat{S}(\omega)-S(\omega)).\end{aligned} \quad (9)$$

It is not difficult to verify that

$$0<\frac{(\lambda^2+4\tilde{S}^2(\omega))^{\frac{1}{2}}\lambda-\lambda^2}{2\tilde{S}^2(\omega)}<1, \quad (10)$$

which proves Proposition 3. □

3.2. Forward Adaptation

In a forward adaptive system, the PSD is estimated from a segment of unquantized signal. Many practical approaches exist to obtain an estimate of the PSD [5]. For example, the autocorrelation method of linear predictive coding (LPC) analysis yields a set of coefficients of an auto-regressive (AR) signal model that can be used to describe the PSD (e.g., by autoregressive spectral estimation). The description of the PSD needs to be transmitted to the decoder. Several methods can be used to quantize such an AR model-based description of the PSD (e.g., by applying quantization in the line spectral frequency domain [6] or by quantizing the spectral immittance pairs [7]). Each segment of the signal is then described by a two-stage description. One stage comprises a quantized model and the other stage comprises a quantized waveform.

The estimation error and quantization of the PSD description lead to a mismatch from the source PSD, which is inevitable in practice. However, according to Proposition 3, the proposed method is relatively robust to such a mismatch. It can also be seen that as λ decreases, the similarity between the reconstructed and the source PSD improves.

3.3. Backward Adaptation

We now consider a backward adaptive system. We divide a stationary process into segments. For each segment, we apply pre/post-filter that are based on the estimation of the source PSD using previously reconstructed samples.

Let $\hat{S}_n(\omega)$ be the PSD estimate, which is used to derive the n-th pre/post-filter $H_n(\omega)$ and $G_n(\omega)$ according to (5) and (6).

The process starts with an initial PSD estimate $\hat{S}_0(\omega)$. In the following, $\tilde{S}_n(\omega)$ denotes the PSD of the reconstruction when the n-th pre/post-filter are used. We are interested in the conditions under which $\tilde{S}_n(\omega)$ approaches the source PSD.

Proposition 3 implies convergence of the backward adaptation. The PSD estimate $\hat{S}_n(\omega)$ is updated from the reconstructed signal and therefore, is related to $\tilde{S}_{n-1}(\omega)$. In the following proposition we see that the PSD of the reconstruction approaches the source PSD if $\hat{S}_n(\omega) = \tilde{S}_{n-1}(\omega)$.

Proposition 4. *If* $\hat{S}_n(\omega) = \tilde{S}_{n-1}(\omega)$ *and* $\hat{S}_0(\omega) > 0, \forall\omega$, *then*

$$\lim_{n\to\infty} \tilde{S}_n(\omega) = S(\omega). \tag{11}$$

Proof. Using $\hat{S}_n(\omega) = \tilde{S}_{n-1}(\omega)$, Proposition 3 implies that $\tilde{S}_n(\omega)$ converges. Taking $\lim_{n\to\infty}$ to both side of (9) and solving for the limit of $\tilde{S}_n(\omega)$, we obtain that $\lim_{n\to\infty} \tilde{S}_n(\omega)$ equals either $S(\omega)$ or zero. However, when $\tilde{S}_0(\omega) > 0$, the latter situation violates Proposition 3 unless $S(\omega) = 0$. Both situations lead to $\lim_{n\to\infty} \tilde{S}_n(\omega) = S(\omega)$ (although $S(\omega) = 0$ is beyond our assumption). □

Proposition 4 also implies that the rate-MSE performance of the backward adaptive PSD-PQ converges to the optimality stated in Proposition 1.

We note that the convergence speed depends on the rate. In particular, observing (9), if λ is large, the factor that governs the convergence is close to 1 and the convergence becomes slow.

4. EVALUATION AND RESULTS

In this section we show that PSD-PQ has observable benefits. We first show that a forward adaptive PSD-PQ performs well for real-world audio coding. We then show that the backward adaptation converges to the optimal rate-MSE performance.

4.1. Application to Audio Coding

As the benefits of PSD-PQ are perceptual in nature, subjective tests are needed for its evaluation. To this purpose we inserted a *predictive* PSD-PQ in a relatively standard platform and compare it to a rate-MSE optimal quantizer.

4.1.1. Audio Coding Platform

We used a forward adaptive predictive audio coder as a platform for our comparison. A diagram of the platform is shown in Fig. 2. The platform consisted of perceptual filters and a predictive quantizer. The perceptual filter and its inverse filter respectively perform transforms to and from the perceptual domain where the MSE criterion can be used [8]. The predictive quantizer is constructed by introducing a predictor into the basic scheme of Fig. 1. An advantage is that the dependency in the source is reduced and thus the entropy coding in the ECDQ does not need any memory. It is shown in [9] that such a predictive quantizer asymptotically achieves the same rate-MSE performance as the scheme in Fig. 1.

We used a short-term AR model (including the gain), extracted from the input signal every 20 ms and interpolated to a 5 ms to describe the spectral envelope of the input signal. The perceptual weighting filters are derived from this short-term AR model. The model estimation and perceptual weighting are similar to AMR-WB [10].

The spectral structure of the signal can be seen as the product of an envelope and a harmonic (fine) structure. The spectral envelope is described by the short-term AR model. Accordingly, the pre/post-filter can be decomposed as a concatenation of an envelope and a pitch filter. To reduce computational complexity in our implementation we neglect the pitch pre-filter, which we found to have relatively little effect. The pitch post-filter is effectively described by an all-zero structure $P(z) = \beta + \gamma z^{-p}$, where β and γ are adjustable parameters and p is the pitch period.

The platform operates on mono 16 kHz sampled audio signals. The model parameters (the AR model and also the pitch) can be transmitted with high fidelity at a rate of about 6 kbps. We adjusted the rate for each input excerpt to obtain the average rate to be $18\pm5\%$ kbps, targeting an overall rate of 24 kbps, which is a common operating rate for communication oriented codecs. This bit rate was also chosen to illustrate the performance of PSD-PQ on a transition between low and high rates. At low rates PSD-PQ has a clear perceptual advantage over MSE-optimized quantization. For example, it is known that the MSE-optimized quantization leads to "birdies" and band limitation at low rates [11]. PSD-PQ does not introduce these artifacts. At high rates, it is expected that the performance of PSD-PQ is equivalent to MSE-optimized quantization [1].

Table 1. Average MUSHRA results$\in [0, 100]$ of the proposed and the reference audio coder for each test signal.

item	content	proposed	reference
es01	English female speaker	76.17	73.83
es02	German male speaker	63.83	59.92
es03	English female speaker	68.50	63.08
sc01	Trumpet solo & orchestra	64.83	71.25
sc02	Symphonic orchestra	80.42	54.00
sc03	Contemporary pop music	78.50	59.00
si01	Harpsichord	56.17	66.08
si02	Castanets	86.42	67.75
si03	Pitch pipe	50.25	47.58
sm01	Bagpipes	47.17	59.42
sm02	Glockenspiel	70.42	75.17
sm03	Plucked strings	78.83	70.83
	Average score	68.46	63.99

4.1.2. Test Systems

To test the PSD-PQ system, we replaced the predictive quantizer of the platform with two different quantizers: our predictive PSD-PQ and the rate-MSE optimized reference predictive quantizer of [9]. The two quantizers are identical in structure. The only difference is the specification of the pre/post-filter.

The reference predictive quantizer [9] is asymptotically rate-MSE optimal for stationary Gaussian processes. It contains pre/post-filters $|H_{\mathrm{ref}}(\omega)|^2 = \lambda^{-1}\left(1 - \frac{\min(S(\omega),\lambda)}{S(\omega)}\right)$ and $G_{\mathrm{ref}}(\omega) = \lambda H^*_{\mathrm{ref}}(\omega)$. These rate-MSE optimized pre/post-filters remove the frequency components that are below a threshold (this corresponds to the "reverse water-filling" phenomenon, which happens when the optimal rate-MSE tradeoff for a stationary Gaussian process is reached without the PSD preserving constraint).

4.1.3. Test Results

We compared the perceptual performance of the proposed and the reference audio coder using the MUSHRA methodology [12]. In the test we used twelve MPEG audio excerpts, resampled to 16 kHz, as the test signals. The coded signals, together with a 3.5 kHz low-passed signal and the unprocessed signal, were presented to listeners in a random order. Twelve listeners participated in the test. The average test results and respective 95% confidence intervals were as follows: unprocessed signal 99.72 $\pm$ 0.43, proposed scheme 68.46 $\pm$ 3.80, reference scheme 63.99 $\pm$3.40, and 3.5 kHz signal 46.06 $\pm$ 3.10. The averaged MUSHRA scores of the proposed and the reference audio coder for each test signal are listed in Tab. 1.

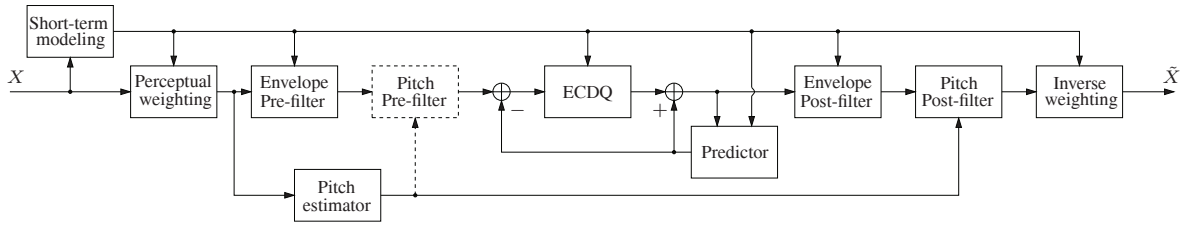

Fig. 2. Diagram of an audio coder platform. The dashed lines indicate negligible components.

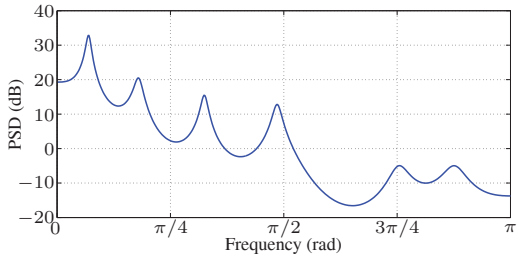

Fig. 3. PSD of the test signal.

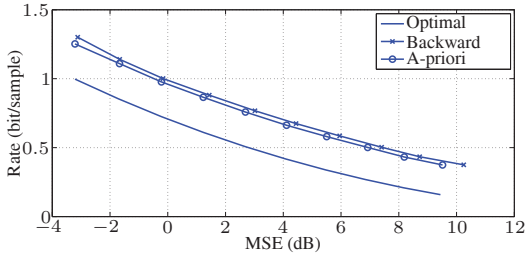

Fig. 4. Simulation results.

The proposed coder is in general better than the reference coder since it preserves the source spectrum while the reference coder removes some frequency bands due to reverse water-filling. The advantage is seen in most items, being most prominent in sc02, sc03, and si02, for which the bandwidth plays a major role in perception. It can also be noticed that the proposed coder performed slightly worse than the reference coder for sc01, si01, sm01, and sm02. These signals have very pronounced pitch and/or multi-pitch structure, which the short-term AR model and single pitch cannot model well. A better signal modeling is expected to enhance the performance.

4.2. Backward Adaptation

To show that the backward adaptation is effective we considered a stationary AR Gaussian process with the PSD shown in Fig. 3. A predictive quantizer with the proposed backward adaptive pre/post-filter was applied. An AR model based estimate of the PSD was updated with an interval of 80 samples using a long asymmetric window that emphasizes recent samples. As a reference we used a PSD-PQ with prior knowledge of the source PSD. The results are shown in Fig. 4, which also includes the optimal rate given in Proposition 1. As long as the $S_0(\omega) > 0, \forall\omega$, the system converges to near-optimal performance. A loss of about 0.254 bits/sample is caused by the use of a scalar quantizer. The backward adaptive system provides a performance that is essentially identical to that of PSD-PQ with prior knowledge of the source PSD. The difference compared to the prior-knowledge case is due to errors in the PSD estimation.

5. CONCLUSIONS

The results of this paper reconfirm that the basic philosophy of distribution preserving quantization (DPQ), to quantize the signal subject to the preservation of statistical properties is an effective method for efficient audio coding across a broad range of rates. The approach operates like a conventional waveform coder at high rates, and the constraint on preserving the statistical properties means that the method operates like parametric coding at low rates.

In this paper we relaxed the DPQ constraint to a power spectral density (PSD) constraint. This facilitates a lower computational complexity and does not impact performance in the context of an audio coder. The PSD constraint can be implemented using forward and backward adaptation. We showed that the PSD constraint can provide better subjective performance than a straightforward minimum mean squared error criterion.

6. REFERENCES

[1] M. Li, "Distribution preserving quantization," Ph.D. dissertation, KTH Royal Institute of Technology, 2011.

[2] M. Li, J. Klejsa, and W. B. Kleijn, "Distribution preserving quantization with dithering and transformation," *IEEE Signal Process. Lett.*, vol. 17, no. 12, pp. 1014–1017, 2010.

[3] R. Zamir and M. Feder, "Information rates of pre/post-filtered dithered quantizers," *IEEE Trans. Inf. Theory*, vol. 42, no. 5, pp. 1340–1353, 1996.

[4] ——, "On lattice quantization noise," *IEEE Trans. Inf. Theory*, vol. 42, no. 4, pp. 1152–1159, 1996.

[5] S. M. Kay, *Modern Spectral Estimation, Theory and Application.* Prentice-Hall, 1988.

[6] P. Kabal and R. Ramachandran, "The computation of line spectral frequencies using Chebyshev polynomials," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 6, pp. 1419 – 1426, 1986.

[7] Y. Bistritz and S. Peller, "Immittance spectral pairs (ISP) for speech encoding," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing ICASSP'93,*, vol. 2, 1993, pp. 9 –12 vol.2.

[8] B. Edler and G. Schuller, "Audio coding using a psychoacoustic pre- and post-filter," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing ICASSP '00*, vol. 2, 2000, pp. 881–884 vol. 2.

[9] R. Zamir, Y. Kochman, and U. Erez, "Achieving the Gaussian rate–distortion function by prediction," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 3354–3364, 2008.

[10] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Jarvinen, "The adaptive multirate wideband speech codec (AMR-WB)," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 8, pp. 620–636, 2002.

[11] C.-M. Liu, H.-W. Hsu, and W.-C. Lee, "Compression artifacts in perceptual audio coding," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 681–695, 2008.

[12] *Recommendation BS.1534-1 — Method for the subjective assessment of intermediate quality level of coding systems*, ITU-R Std., 2003.