TRANSFORM AUDIO CODING WITH ARITHMETIC-CODED SCALAR QUANTIZATION AND MODEL-BASED BIT ALLOCATION

Marie Oger[1] *, Stéphane Ragot*[1] *and Marc Antonini*[2]

[1]France Télécom R&D/TECH/SSTP, Av. Pierre Marzin, 22307 Lannion Cedex
[2]Lab. I3S-UMR 6070 CNRS and Univ. of Nice Sophia Antipolis, rte des Lucioles, 06903 Sophia Antipolis
E-mail: {marie.oger, stephane.ragot}@orange-ftgroup.com, am@i3s.unice.fr

ABSTRACT

In this paper we present a new model-based method to code the transform coefficients of audio signals. The histogram of transform coefficients is approximated by a generalized Gaussian model for efficient model-based bit allocation and the spectrum is coded by scalar quantization followed by arithmetic coding. An example coder operating at 16 kHz and using predictive modified discrete cosine transform (MDCT) coding is described. We compare the performance of the proposed coder with ITU-T G.722.1. Objective and subjective quality results are presented. The proposed coder is better than ITU-T G.722.1 at 24 kbit/s and equivalent at 32 kbit/s.

Index Terms— Transform coding, audio coding, modeling.

1. INTRODUCTION

The telecommunication market is currently evolving towards new multimedia services over IP. In particular, many efforts are focused on improving audio quality by moving from narrowband speech coding (300-3400 Hz) to wideband speech coding (50-7000 Hz). Besides extending audio bandwidth, quality can be improved by optimizing the quantization of existing coding models. ITU-T G.722.1 Recommendation [1] is an example of wideband speech and audio coding system that is used in audio and video conferencing applications. This coder is built upon modified discrete cosine transform (MDCT), scalar quantization and vector Huffman coding of normalized MDCT coefficients. We propose in this work a different coding method for MDCT coefficients, with the objective to improve coding efficiency.

The main contribution of this work lies in the use of arithmetic-coded scalar quantization and the application of generalized Gaussian modeling for efficient bit allocation. Generalized Gaussian modeling is commonly used in image and video coding [2] but its application to speech and audio coding is quite new.

This paper is organized as follows. We present the generalized Gaussian model in Section 2, we give an modeling example for normalized MDCT coefficients. Quantization with model-based allocation and arithmetic coding is described in Section 3. An example coder using the proposed quantization is given in Section 4. Objective and subjective quality results are presented and discussed in Section 5 before concluding in Section 6.

This work was supported in part by the European Union under Grant FP6-2002-IST-C 020023-2 FlexCode.

2. GENERALIZED GAUSSIAN MODELING

2.1. Definition of the generalized Gaussian pdf

The probability density function (pdf) of a zero-mean generalized Gaussian random variable z of standard deviation σ is given by [2]:

$$g_\alpha(z) = \frac{A(\alpha)}{\sigma}\, e^{-|B(\alpha)z/\sigma|^\alpha}, \tag{1}$$

where α is a shape parameter describing the exponential rate of decay and the tail of the density function. The parameters $A(\alpha)$ and $B(\alpha)$ are given by:

$$A(\alpha) = \frac{\alpha B(\alpha)}{2\Gamma(1/\alpha)} \qquad \text{and} \qquad B(\alpha) = \sqrt{\frac{\Gamma(3/\alpha)}{\Gamma(1/\alpha)}}, \tag{2}$$

where $\Gamma(.)$ is the Gamma function defined as:

$$\Gamma(\alpha) = \int_0^\infty e^{-t} t^{\alpha+1}\, dt. \tag{3}$$

The Laplacian and Gaussian distributions correspond to the special case $\alpha = 1$ and 2 respectively. The generalized Gaussian model is useful to approximate symetric unimodal distributions.

2.2. Estimation of the shape parameter α

Methods to estimate the shape parameter α of a generalized Gaussian random variable are reviewed in [3]. We use hereafter the open-loop method proposed by Mallat to estimate α. For a generalized Gaussian variable z, a relation between the variance $E(z^2)$, the mean absolute value $E(|z|)$ and the shape parameter α is given by [4]:

$$\frac{E(|z|)}{\sqrt{E(z^2)}} = \frac{\Gamma(2/\alpha)}{\sqrt{\Gamma(1/\alpha)\Gamma(3/\alpha)}} = F(\alpha) \tag{4}$$

The shape parameter α can be therefore estimated as:

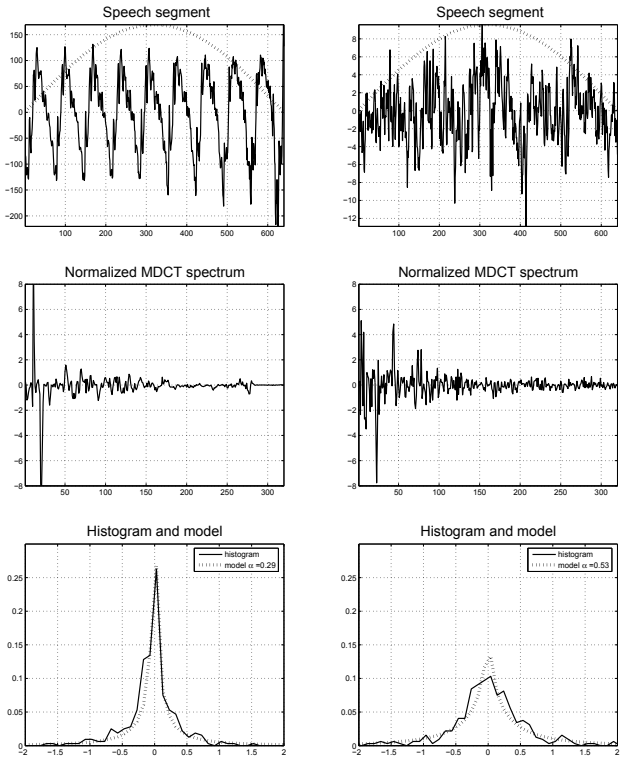
$$\hat{\alpha} = F^{-1}\left(\frac{\hat{m}_1}{\sqrt{\hat{m}_2}}\right) \tag{5}$$

where $\hat{m}_1 = \frac{1}{n}\sum_{i=1}^n z_i^2$ and $\hat{m}_2 = \frac{1}{n}\sum_{i=1}^n |z_i|$ are measured on available data $\{z_1, \ldots, z_n\}$.

2.3. Estimation examples for speech

Figure 1 shows how the distribution of normalized MDCT coefficients can be approximated by a generalized Gaussian model. Both

voiced and unvoiced speech examples are considered. The input signals in time domain (top) are sampled at 16 kHz and transformed by MDCT with a sinusoidal window of 40 ms. The MDCT spectrum of a given frame (middle), comprising 320 coefficients, is normalized by its root mean square (r.m.s.). The histogram of normalized MDCT coefficients is modeled by a generalized Gaussian pdf (bottom). The estimated shape parameter for voiced speech is $\alpha = 0.29$ and for unvoiced speech it is $\alpha = 0.53$, so the value of α is somehow related to the voicing of the signal. Note that the value of α would be closer to 2 (Gaussian case) for unvoiced speech if the signals in time domain were linear predictive residuals. However these examples indicate that generalized Gaussian modeling can provide a good approximation of the MDCT spectrum distribution for speech signals. These observations can be extended to music signals.



(a) Voiced segment (b) Unvoiced segment

Fig. 1. Examples of MDCT coefficient modeling.

3. ENTROPY-CODED SCALAR QUANTIZATION WITH MODEL-BASED ALLOCATION

We follow here the notations of [5] with regards to transform coding and bit allocation. We consider the encoding of N zero-mean random variables $x_1, \ldots, x_N$ of variances $\sigma_1^2, \ldots, \sigma_N^2 > 0$ with respect to the mean square error criterion. The variables x_i are coded by scalar quantization with the *same* step size q. We assume that the sequence of integers obtained after scalar quantization is encoded by ideal entropy coding.

In case of high resolution the mean square error D resulting for the encoding of N random variables x_i is given by [5]:

$$D \approx \sum_{i=1}^{N} h_i \sigma_i^2 2^{-2b_i} \tag{6}$$

where the constant h_i is a function of the pdf of the variable x_i and b_i is the number of bits per sample used to code x_i. We assume that the variables x_i have a generalized Gaussian pdf, in this case h_i is given by [2]:

$$h_i = \frac{\Gamma(1/\alpha_i)^3}{3\alpha_i^2 \Gamma(3/\alpha_i)} e^{2/\alpha_i} \tag{7}$$

where α_i is the shape parameter of x_i. For a given bit allocation B in bits per sample, the bit allocation problem is to minimize the distortion D under the constraint that $\sum_{i=1}^{N} b_i \leq B$. The distortion D given in Eq (6) can be minimized with Lagrangian techniques. The criterion $J(b_i, \lambda)$ is defined as

$$J(b_i, \lambda) = D - \lambda \left(\sum_{i=1}^{N} b_i - B \right) \tag{8}$$

where λ is the Lagrange multiplier. It can be shown that the optimal λ is given by [2]:

$$\lambda_{opt} = 2 \ln(2) \sum_{i=1}^{N} \sigma_i^2 h_i 2^{-2b_i} \tag{9}$$

With this value of λ the distortion D becomes:

$$D = \frac{\lambda_{opt}}{2 \ln(2)} \tag{10}$$

Furthermore for high-resolution scalar uniform quantization with step size q, we have [5]:

$$D = \frac{q^2}{12} \tag{11}$$

From (10) and (11) we find that the optimal stepsize is:

$$q = \sqrt{\frac{6\lambda_{opt}}{\ln(2)}} \tag{12}$$

In this work, a single shape parameter α is estimated (i.e. $\alpha_i = \alpha$) and the entropy coding is implemented using stack-run coding [6], which has been originally developed for image coding. Stack-run coding represents a sequence of signed integers by adaptive arithmetic coding using a quaternary alphabet (0, 1, +, -) and switched contexts. This entropy coder implies a biais in bit consumption.

To verify the fixed bit allocation constraint the step size q is determined in practice in two steps:

1. Estimation of optimal step size q as in Eq (12).
2. Step size refinement by bisection search to verify the bit allocation constraint $\sum_{i=1}^{N} b_i \leq B$.

4. EXAMPLE CODER USING THE PROPOSED METHOD

4.1. Encoder

The proposed coder is illustrated in Figure 2. The encoder employs a linear-predictive weighting filter followed by MDCT coding. The input sampling frequency is 16000 Hz, while the frame length is 20 ms with a lookahead of 25 ms. The effective bandwidth of the input signal is considered to be 50-7000 Hz. An elliptic high-pass

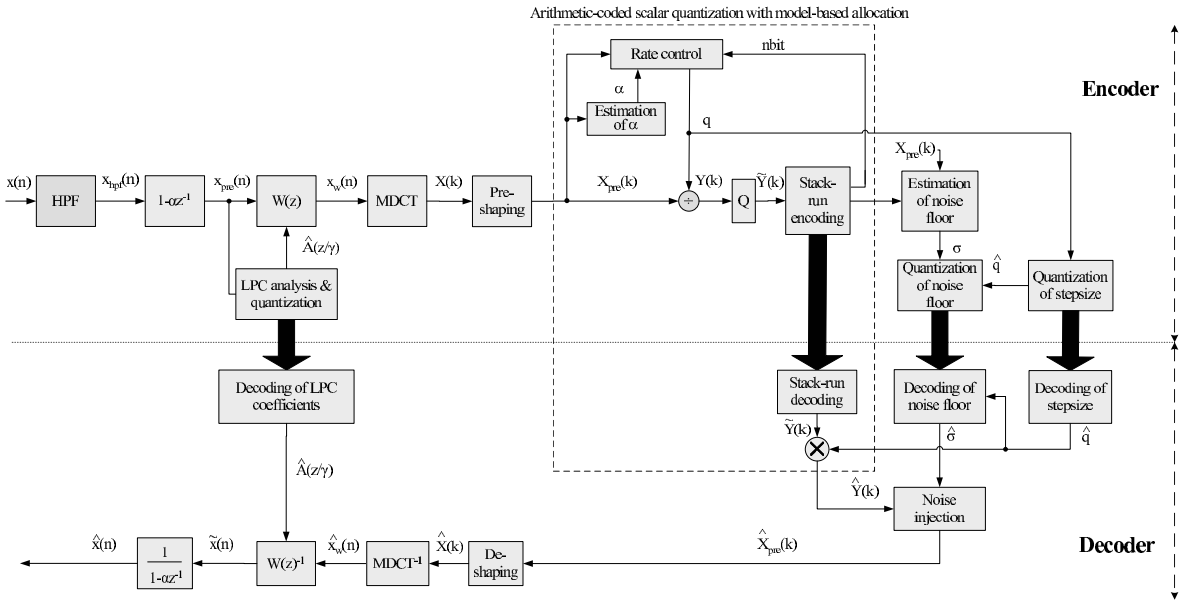


Fig. 2. Block diagram of the proposed predictive transform coder.

filter (HPF) is applied to the input signal $x(n)$ in order to remove the frequency component under 50 Hz. The resulting signal $x_{hpf}(n)$ is then preemphasized by $1 - \alpha z^{-1}$ with $\alpha = 0.75$. An 18th order LPC analysis described in [7] is performed on the preemphasized signal $x_{pre}(n)$. The resulting LPC coefficients are quantized with 40 bits using a parametric method based on a Gaussian mixture model (GMM) [8] in the linear spectrum frequency (LSF) domain. The pre-emphasized signal is filtered by a perceptual weighting filter:

$$W(z) = \frac{\hat{A}(z/\gamma)}{1 - \beta z^{-1}} \tag{13}$$

where $\beta = 0.75$ is a tilt parameter and $\gamma = 0.92$. The coefficients of $W(z)$ are updated every 5 ms by interpolating LSF parameters. An MDCT analysis is applied on the weighted signal $x_w(n)$. The MDCT is implemented using the fast algorithm of [9] which is based on a complex FFT. Then the MDCT coefficients $X(k)$ are pre-shaped to emphasize low frequencies, in a way similar to 3GPP AMR-WB+ [10]. The pre-shaped coefficients $X_{pre}(k)$ are divided by a step size q and the resulting spectrum $Y(k)$ is encoded by scalar quantization. For a given spectrum $Y(k)$ the spectrum $\tilde{Y}(k)$ after scalar quantization is defined as:

$$\tilde{Y}(k) = [Y(k)] = \left[\frac{X_{pre}(k)}{q}\right] \tag{14}$$

where [.] represents the rounding to the nearest integer. Only the first 280 coefficients of the $Y(k)$ spectrum corresponding to the 0-7000 Hz band are coded; the last 40 coefficients are discarded. The integer sequence $\tilde{Y}(k)$ is encoded by stack-run coding [6]. The rate control consists in finding the appropriate step size q so that the number of bits, $nbit$, used for stack-run coding matches the allocated bit budget as described in Section 3. The distribution of $X_{pre}(k)$ is approximated by a generalized Gaussian model and the shape parameter α is estimated using Mallat's method. Then a noise level estimation is performed on the spectrum $Y(k)$ after stack-run coding. The noise floor σ is estimated as:

$$\sigma = r.m.s.\,\{X_{pre}(k)\,|\,Y(k) = 0\} \tag{15}$$

with the additional constraint that $Y(k)$ must belong to a long zero run to be really considered in the above r.m.s. calculation. The step size q is scalar quantized in log domain with 7 bits. The noise floor σ is quantized by coding the ratio $\sigma/\hat{q}$ in linear domain with 3 bits.

4.2. Decoder

The decoded LSF are interpolated every 5 ms and converted to LPC coefficients. The reconstructed spectrum $\hat{Y}(k)$ is given by:

$$\hat{Y}(k) = \hat{q}\ \tilde{Y}(k) \tag{16}$$

where $\tilde{Y}(k)$ is found by stack-run decoding. In order to improve quality, noise injection is applied on $\hat{Y}(k)$. A noise of magnitude $\pm\hat{\sigma}$ is injected in all zero sequences longer than 20 coefficients in $\hat{Y}(k)$. The spectrum $\hat{X}(k)$ is de-shaped in a way similar to 3GPP AMR-WB+ and transformed in time domain using the inverse MDCT and overlap-add algorithm described in [9]. An inverse perceptual filter $W(z)^{-1}$ is applied on $\hat{x}_w(n)$ in order to shape the coding noise introduced in the MDCT domain. The response of $W(z)^{-1}$ is similar to a short-term masking curve and its coefficients are updated every 5 ms by LSF interpolation. The signal $\tilde{x}(n)$ is deemphasized to find the synthesis $\hat{x}(n)$.

4.3. Bit allocation

The parameters of the proposed coder are the Line Spectrum Frequency (LSF) parameters, the step size, and the noise floor level. The bit allocation to the parameters is detailed in Table 1, where B_{tot} is the total number of bits per frame. For instance at 24 kbit/s, $B_{tot} = 480$ bits. The allocation (in bits per sample) to stack-run coding is $B = (B_{tot} - 50)/280$.

Table 1. Bit allocation for the coding scheme.

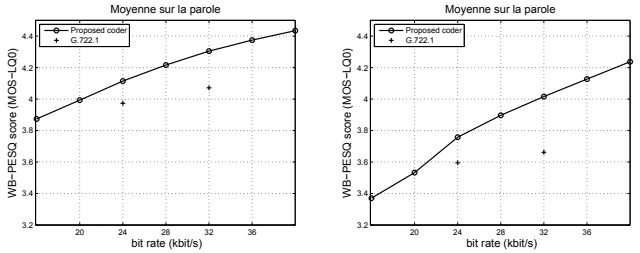
Parameter	Number of bits
LSF	40
Step size q	7
Noise floor σ	3
Stack-run coding	B_{tot}-50
Total	B_{tot}

5. EXPERIMENTAL RESULTS

A database of 24 clean speech samples in French language (6 male and female speakers×4 sentence-pairs) and 16 clean music samples (4 types×4 samples) of 8 seconds is used. These samples are sampled at 16 kHz, preprocessed by the P.341 filter of ITU-T G.191A and normalized to -26 dB_{ov} using the P.56 speech voltmeter.

5.1. Objective quality results

WB-PESQ [11] is used to evaluate the quality of the proposed coder and compare it with ITU-T G.722.1. Only clean speech samples are used to compute the average WB-PESQ scores at various bitrates. The bit rate varies from 16 to 40 kbit/s with a step of 4 kbit/s for our coder. ITU-T G722.1 is tested at 24 and 32 kbit/s. Figure 3 shows the WB-PESQ scores obtained for the two coders, by considering separately male and female cases. These results suggest that the quality of the proposed coder is better than ITU-T G.722.1 at 24 and 32 kbit/s (0.2-0.3 MOS-LQ0 difference). Note that these results show a clear male/female dependency, which was already observed in formal ITU-T G.722.1 subjective test results [12].



(a) Male (12 sentence-pairs) (b) Female (12 sentence-pairs)

Fig. 3. Average WB-PESQ score.

5.2. Subjective quality results

Two informal AB tests at 24 kbit/s have been conducted: one for speech, another for music. In total 8 experts participated in the test. Figure 4 shows the results. The proposed coder was preferred for music in 53% of cases and for speech in 48% of cases. The results confirmed the objective quality results at 24 kbit/s. Subjective tests have also been conducted at 32 kbit/s but the quality improvement of the proposed coder is less significant and the two coders are equivalent. The proposed coder is better than G.722.1 at 24 kbit/s and equivalent at 32 kbit/s, but it has a higher complexity than G.722.1.

The discrepancy between objective and subjective results at 24 and 32 kbit/s can be explained by the large sensitivity of WB-PESQ in low frequencies – indeed the proposed coder has in general a lower distortion in low frequencies –, and by the limited applicability of WB-PESQ to compare distortions of different natures.

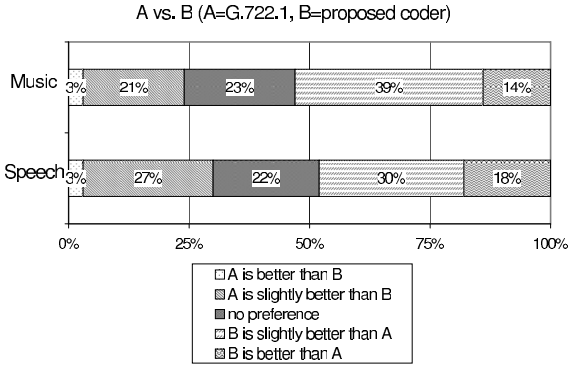


Fig. 4. AB test results at 24 kbit/s.

6. CONCLUSION

In this paper we proposed an MDCT coder with generalized Gaussian modeling for wideband speech and audio signals sampled at 16 kHz. This coder was compared with ITU-T G.722.1. The quality improvement is mainly due to the use of arithmetic coding (instead of Huffman coding) and perceptual filtering. The generalized Gaussian model allows to minimize the complexity of bit allocation by estimating efficiently the quantization stepsize. The proposed spectrum coding technique could be applied to code FFT coefficients in order to improve the quality of TCX modes of 3GPP AMR-WB+.

REFERENCES

[1] ITU-T G.722.1, *Coding at 24 kbit/s and 32 kbit/s for Hand-free Operations in Systems with Low Frame Loss*, 1999.

[2] C. Parisot, M. Antonini, and M. Barlaud, "3d scan based wavelet transform and quality control for video coding," *EURASIP*, vol. 2003, no. 1, pp. 521–528, Jan 2003.

[3] M. Oger, S. Ragot, and M. Antonini, "Low-complexity wideband LSF quantization by predictive KLT coding and generalized gaussian modeling," *in Proc. Eusipco*, 2006.

[4] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 11, pp. 674–693, July 1989.

[5] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer, 1993.

[6] M. J. Tsai and al., "Stack-run image coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, no. 5, pp. 519–521, Oct. 1996.

[7] M. Oger, J. Bensa, S. Ragot, and M. Antonini, "Stack-run audio coding," *in 120th AES convention*, May 2006.

[8] A. D. Subramaniam and B. D. Rao, "Pdf optimized parametric vector quantization of speech line spectral frequencies," *IEEE Trans. Speech and Audio Proc.*, vol. 11, no. 2, pp. 130–142, Mar. 2003.

[9] P. Duhamel and al., "A fast algorithm for the implementation of filter bank based on time domain aliasing cancellation," *in Proc. ICASSP*, pp. 2209–2212, May 1991.

[10] 3GPP TS 26.290, "Audio codec processing functions; extended adaptive multi-rate - wideband (AMR-WB+) codec; transcoding functions," 2005.

[11] ITU-T Rec P.862.2, *Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs*, Nov 2005.

[12] "Wideband Qualification Test Results for the PictureTel Transform Codec (PTC)," Contribution D.105, ITU-T Q.22/16, Study Period 1997-2000, Geneva, (Source: PictureTel).