LOW DELAY CODING OF WIDEBAND AUDIO (20 HZ - 15 KHZ) AT 64 KBPS

A. Jbira, N. Moreau

ENST 46 rue Barrault
75634 Paris Cedex 13 France
e-mail : jbira@sig.enst.fr, moreau@sig.enst.fr

P. Dymarski

Technical University of Warsaw
00-665 Warsaw Poland
e-mail : dymarski@tele.pw.edu.pl

ABSTRACT

A 64 kbps coder of wideband (15 kHz) monophonic audio signals is described. Its structure is based on the transform coded excitation scheme, adopted to 7 kHz band signals. Significant modifications are proposed, that yield the reduction of delay while keeping an almost transparent quality of speech and music, equivalent to that provided by the MPEG1, layer II audio standard at the same bit rate. Algorithmic delay has been reduced to 17 ms - approximately 1/3 the delay of the MPEG coder.

1. INTRODUCTION

The work presented in this paper is realized in the framework of studies aiming to develop new services for group communication (audio-video teleconferencing, telephony on loud-speaker, etc.). The existing ITU-T standard G722 offers 7 kHz bandwidth and acceptable quality at 64 kbps. In order to meet the growing bandwidth and quality demands, the ITU-T is currently standardizing 7 kHz speech coding at 16, 24 and 32 kbps. However the high quality teleconferencing requires wider bandwidth (up to 16 kHz), good quality for speech and music at maximum 64 kbps and reduced delay.

A wideband audio coder (sampling frequency Fs = 32 kHz), susceptible to transmit music signals as well as speech signals at bit rate 64 kbps or lower per monophonic channel has been developed. The 64 kbps coder only is described here. The originality of this work comes from the priority given to minimizing the delay while achieving near transparent quality equivalent to that provided by the MPEG-1, layer II audio standard at the same bit rate. Indeed, while the delay constraint is weaker for diffusion type applications, it is stronger for applications dealing with group communication. Therefore, the maximal admissible algorithmic delay has been fixed to 20 ms (approximately 3 times lower the delay allowed by MPEG-audio standard).

According to the nature of the processed signal, audio coders can be classified into two main categories : speech and music coders. Generally speech coders use a predictive vector quantization. The most typical example of such coders is the CELP coder. These coders have the advantage of being based on analysis-by-synthesis scheme allowing a closed-loop quantization. The auditory model used, based on the whitening filter, is simplified because at low bit rates and in telephonic band, a sophisticated auditory model is not necessary. Speech coders yield small reconstruction delays, typically 10-20 ms. However, they are less adapted to high bit rates ; the computational complexity becomes a serious problem over 20 kbps. The second category of coders, music coders, uses a time-frequency decomposition to exploit an elaborate psychoacoustic model. Bits are allocated per transform component in order to mask the reconstruction noise. A time-frequency decomposition (using a filter bank or a transform) is first applied to the signal samples and then transform components are encoded and transmitted. The main constraint of such coders is that they need a high frequency resolution to make the auditory model efficient. However, this requires long analysis windows which results in high reconstruction delays.

The coding scheme represented in Fig. 1, realizes a compromise between the two preceding categories. It has been subject to several studies for telephonic band, for example the OTC coder introduced in [1], and for 7 kHz band, for example the TCX [2] and TPC [3] coders. In this scheme, the time varying analysis filter $W(z)$, based on the whitening filter, is used as in CELP coders to transform the signal to the so-called perceptual space. However, the quantization of the perceptual signal is realized, as in music coders,

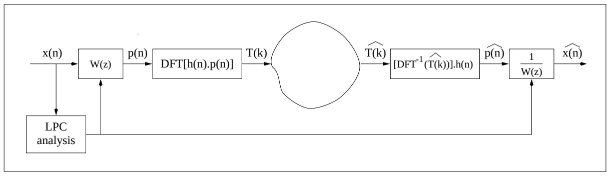

Figure 1: Principle of a “mixed” coding scheme realizing a compromise between “CELP coders” and “frequency coders”.

This work was supported by France Telecom - CNET

using a time-frequency decomposition. This quantization is done under the control of a bit allocation procedure in order to make the noise spectrum flat. The proper spectral shape of the reconstruction noise is obtained by the synthesis filter $1/W(z)$. For TCX coder as well as for TPC coder, the first time-frequency decomposition used was the Discrete Fourier Transform (DFT) without overlap. Other transforms have been investigated in [4] and [5].

In [2] the transform coded excitation (TCX) scheme, shown in Fig. 1, has been used for coding of music signals in 7 kHz band at 24 kbps. In this paper we describe a series of modifications, necessary to adopt this scheme for coding of speech and music signals in 15 kHz band at 64 kbps. Preliminary subjective tests show that the requested quality constraint is satisfied while maintaining a low algorithmic delay (17 ms).

2. DESCRIPTION OF THE CODER

The four steps of coding are described successively : the construction of the perceptual signal, the time-frequency decomposition, the amplitude and phase spectrum quantization.

2.1. Construction of the perceptual signal

Generally a perceptual filter of the form

$$W(z) = (1 - \mu z^{-1}) \frac{A(z/\gamma_1)}{A(z/\gamma_2)}$$

is used. $A(z)$ is the whitening filter obtained by a standard LPC analysis. A "split VQ" quantization of LSP coefficients is applied. A linear interpolation between LSP coefficients of adjacent analysis windows is realized in order to calculate the perceptual signal $p(n)$ in the whole analysis window.

For our wideband audio coder, we have kept frames of the same length as for the 7 kHz band TCX coder (16 ms). Hence, $N = 512$ samples per frame. The main modifications concern the order of the filter and the quantization of the coefficients. We have chosen a high filter order $P = 50$ instead of $P = 16$. Indeed, evaluation of the prediction gain shows that this gain grows significantly with P, for audio and speech signals at a sampling rate of 32 kHz, and saturates at $P = 50$. On the other hand, the suppression of the pitch predictor requires the increase of the order as in the LD-CELP coder. The total of coefficients are coded by using 90 bits (9% of the bit rate). A linear interpolation is realized on LSP coefficients every 2 ms.

2.2. Time-frequency decomposition

The properties of signals to be coded (speech and music) have an important consequence on choosing the characteristics of the time-frequency decomposition. To code speech signal, a pitch predictor is very helpful, especially in telephonic band. However, it requires short analysis windows and this highly constrains on frequency resolution of the time-frequency decomposition. When it comes to music signals, the pitch predictor is less successful. This is because music signals are not exactly harmonic. Therefore, in the music version of TCX, the pitch predictor is not used.

For wideband (15 kHz) audio signals, the problem must be seen in a different way. The harmonicity of speech signals that we exploit using a long term prediction, occurs at maximum up to a frequency of 2 kHz. Moreover, the demand of compressed audio quality is more important for music signals. Thus we have eliminated the long term predictor in order to obtain a high frequency resolution in the time-frequency decomposition. An overlap between the analysis windows is introduced because it does not not only improve the spectral characteristics of the transform, but also attenuates some auditive parasite phenomena caused by block effects. This overlap must remain small in order to prevent increase of the algorithmic delay. On the other hand, if the transformation is not computed with maximal decimation, the overlap should not be penalized in terms of bit rate.

We have decided to introduce a weighting window $h(n)$, with (small) overlap, followed by a DFT. The weighting window must guarantee the cancellation of overlapped factors : without any quantization the reconstruction of original signal must be perfect (or almost perfect). The chosen window is
$h(n) = sin((0.5 + n)\pi/2\tau)$ if $0 \leq n \leq \tau - 1$,
$h(n) = 1$ if $\tau \leq n \leq N - \tau - 1$ and
$h(n) = cos((0.5 + n - N + \tau)\pi/2\tau)$ if $N - \tau \leq n \leq N - 1$.
We choose $\tau = 32$ so with $N = 512$ we have an algorithmic delay of (16 + 1) ms and a "rate cost" of 6%. Fig. 2 shows the frequency representation of the proposed window in comparison with the Hann window and the rectangular window of the same dimension. The output of this transform is a complex vector $[T(0) \cdots T(Np - 1)]$ of dimension $Np = 544$ that we will call spectrum. The magnitude of this vector may be interpreted as the square root of the periodogram.

2.3. Amplitude spectrum quantization

Among the originalities of the TCX coder [2], it realizes a first order prediction on the amplitude spectrum. The optimal prediction coefficient is calculated using the amplitude spectrum in the actual window and the quantized one in the last window. After quantization of this coefficient, the vector of residual amplitude is divided into M sub-vectors of the same dimension. For each of these sub-vectors, a simple gain-shape vector quantization is performed. Then the quantized amplitude spectrum is calculated. This quan-

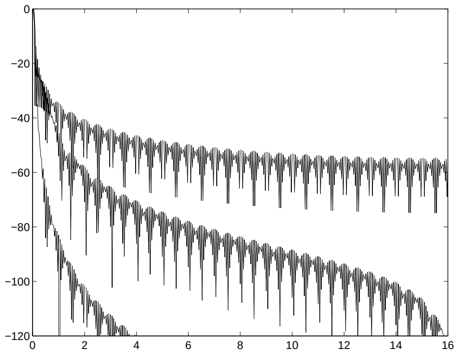

Figure 2: Frequency representation of the rectangular window, the proposed window and the Hann window of the same dimension (from up to down).

tized information is used to normalize all transform components. In this manner, the quantization errors due to amplitude quantization block may be compensated for in phase spectrum quantization block.

Several modifications have been made in the procedure of amplitude spectrum quantization to adapt it to the 15 kHz band audio signals.

The amplitude vector is divided into M sub-vectors of unequal dimension, in order to increase the quantizer performance in low frequencies. Two strategies were tested : the first one consists in splitting the spectrum according to the Bark scale. The second one consists in constructing sub-vectors in such a way that the mean number of bits allocated per sub-vector is the same for all sub-vectors. The frequency partitioning is calculated as follows : for each analysis window of $Np = 544$ samples, a standard bit allocation is calculated using a non quantized amplitude spectrum. A mean allocation of bits is then used to calculate dimensions of M sub-vectors. In Fig.3, the mean allocation of bits and $M = 13$ subbands are shown.

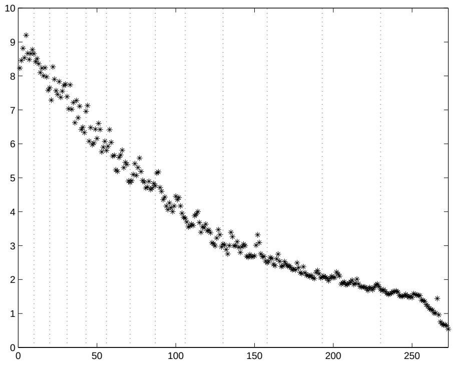

Figure 3: Mean allocation of bits per transform components (simulations were done on 40 s of different speech and music files).

A “mean-removed gain-shape split VQ” is then used to quantize the residual amplitude spectrum. Mean levels are calculated and quantized for each sub-vector before realizing a gain-shape quantization. This is justified first by the fact that, for some sub-vectors, the knowledge of the quantized mean levels to which we add the predicted information are sufficient to have a good reconstruction of the amplitude. Another justification comes from this observation : when the prediction of the amplitude spectrum is less efficient (in case of transitions), the components of the sub-vectors have positive values and should be centred before quantization. We begin by determining the M mean values. We determine the most powerful mean and quantify it on 8 bits according to a logarithmic law. Each average is then divided by the quantized maximal value and coded with 5 bits according to an uniform law.

As we don’t try to do a gain-shape vector quantization for every sub-vector, we consider the problem of selection of the $M' < M$ sub-vectors. These M' sub-vectors do not have to be inevitably the M' first sub-vectors, they are rather the M' vectors of maximum magnitude. Several strategies using an available information at the decoder that avoid a side information to be transmitted, have been tested. This selection can be based on a deduced information from coefficients of $W(z)$. It can also be based on the amplitude spectrum approximation, obtained by a sum of predicted amplitude spectrum and quantized M mean values of residual amplitude spectrum.

The principle of the quantization of the amplitude spectrum is illustrated in Fig. 4. We choose $M' = 5$ sub-vectors among $M = 13$. For each of the selected vectors we attribute 11 bits : 7 for the shape and 4 for the gain. The whole amplitude spectrum quantization requires 143 bits (14% of bit rate).

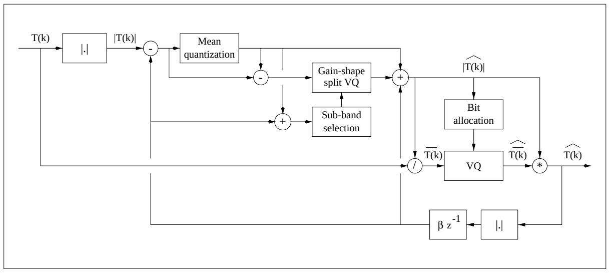

Figure 4: Principle of the quantization of $|T(k)|$.

2.4. Phase spectrum quantization

If the amplitude quantization was perfect, $T(k)$ divided by $|\hat{T}(k)|$ would represent the argument of a unit complex number and a simple uniform scalar quantization would be sufficient. However, because of a small number of bits attributed to amplitude quantization, it is necessary to compensate for the amplitude quantization errors at the phase quantization stage. The solution adopted in the 7 kHz band TCX coder has been to construct all codebooks using training sets, each codebook being composed by 2^b vectors of dimension 2 with b varying from 2 to 8.

For our wideband audio coder, all codebooks are also constructed by training. Our investigation has shown, that different codebooks are required for each of the frequency subbands. In order to obtain training sets, we have selectioned values $T(k)/|\hat{T}(k)|$ for which the number of allocated bits was greater than zero and gathered them according to the partitioning of the frequency axis defined in the last section. A visualization in the complex plane (not reproduced here) shows that the phase components have an uniform distribution between 0 and 2π. On the other hand, the distribution of the magnitude components varies with frequency. Fig. 5 shows an estimation of the probability density function of the magnitude components. The solid lines are relative to 2 first sub-bands, dashed lines to 2 following and dotted lines to 9 last. Fig. 6 shows two codebooks with 256 reproduction values when the number of subband is $k = 1$ (left) or $k = 13$ (right).

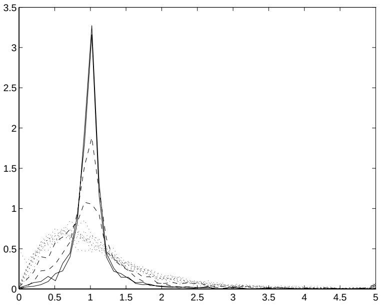

Figure 5: Estimation of the probability density function of magnitude components.

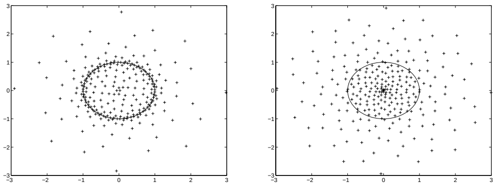

Figure 6: Codebooks used to quantize phase spectrum components when $b = 8$ for “low frequencies” (at left) and for “high frequencies” (at right).

We have investigated the bit allocation block in some detail. *A priori*, the bit allocation has not for objective in this type of coder to give a spectral shape for the reconstructed noise (this is accomplished by the synthesis filter $1/W(z)$). So the remaining bits after quantization of the LPC predictor parameters and amplitude spectrum, have to be allocated according to the distribution of the power spectrum of the target signal - so as to flatten the power spectrum of the quantization error. However the psychoacoustical model offered by the synthesis filter is too much simplified and necessary corrections have to be made at the bit allocation stage. First of all, according to the MPEG psychoacoustical model, the high frequency quantization noise is better masked than the low frequency noise. To take this into consideration, we privilege low frequencies in our bit allocation algorithm.

3. CONCLUSION

The transform coded excitation approach has been used, with a series of necessary modifications, to obtain a wideband 64 kbps low-delay audio coder. These modifications concern the application of an extended weighting window, a mean removed shape gain split VQ for quantization of the amplitude spectrum and multiple codebooks for quantization of the phase spectrum. According to the informal listening tests our coder offers the auditive quality for speech and music, equivalent to that provided by the MPEG-1, layer II audio standard at the same bit rate, but with a delay of 17 ms, which is almost three times lower.

4. REFERENCES

[1] N. Moreau and P. Dymarski, “Successive orthogonalizations in the multistage CELP coder,” *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pp. 61–64, 1992.

[2] R. Lefebvre, R. Salami, C. Laflamme, and J. Adoul, “High quality coding of wideband audio signals using transform coded excitation (TCX),” *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pp. 193–196, 1994.

[3] J. Chen and D. Wang, “Transform predictive coding of wideband speech signals,” *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pp. 275–278, 1996.

[4] J. LeRoux, R. Lefebvre, and J. Adoul, “Comparison of the wavelet decomposition and the Fourier transform in TCX encoding of wideband speech and audio,” *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pp. 3083–3086, 1995.

[5] J. Chen, “A candidate coder for the ITU-T’s new wideband speech coding standard,” *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pp. 1359–1362, 1997.