VERY LOW DELAY AND HIGH QUALITY CODING OF 20 HZ - 15 KHZ SPEECH SIGNALS AT 64 KBIT/S

C. Murgia[1], G. Feng[1], A. Le Guyader[2] & C. Quinquis[2]
murgia@icp.grenet.fr

[1] Institut de la Communication Parlée, URA CNRS n° 368, INPG/ENSERG,
Université Stendhal, GRENOBLE, FRANCE
[2] CNET Lannion A, TSS/CMC, 22301 LANNION, FRANCE

ABSTRACT

In this paper, an algorithm for coding 20 Hz - 15 kHz speech signals at 64 kbit/s with a very low delay (frame of 0.16 ms) is presented. To achieve a quality near to transparency, we propose adapting the Low-Delay CELP coder [1] to the 15 kHz bandwidth and suggest a new noise shaping method based on a psycho-acoustic model. In this way we take advantage of linear predictive coding and masking properties of the human perception system. Finally, an algebraic codebook is proposed, allowing an important reduction of coder computational complexity, without decreasing the perceived quality of signals.

1. INTRODUCTION

High quality speech coding plays an important role in modern telecommunication systems, especially in videoconference and teleconference systems. Compared with narrow band (300 Hz - 3400 Hz), or wideband (50 Hz - 7 kHz) speech, the expansion to the 20 Hz - 15 kHz bandwidth increases significantly the quality of the signals, and improves the sensation of remote speaker presence. To ensure good interactivity in the communication systems, the one-way delay should not exceed 100 ms including processing, sound recording and transmission delays, leaving only about 20 ms for the encoder-decoder delay. To achieve these aims, we propose a high quality coding algorithm at 64 kbit/s for the 20 Hz - 15 kHz speech signal. The starting point of our research was the Low-Delay Code-Excited Linear Prediction method (LD-CELP), originally developed for narrow band signals [1]. Several problems (filter stability, coding noise, etc.) arise from the extension of this algorithm to the 20 Hz - 15 kHz bandwidth [2]. In this paper, we suggest a new noise-shaping method based on a psycho-acoustic model of the human perceptual system. Moreover, in order to reduce the complexity of this algorithm, we propose the use of an ternary algebraic codebook.

2. SYSTEM OVERVIEW

The proposed coding system is based on backward adaptive CELP coding [1]. The basic principles of this algorithm are summarised as follows. The LPC coefficients are updated at the encoder (local decoder) and at the decoder by backward adaptive linear prediction on the previously synthesised signal. Only the excitation parameters are transmitted to the decoder. The excitation signal is a 5-sample vector, selected from a 7 bits shape codebook using the analysis by synthesis method and scaled by a gain factor quantized on 3 bits. With this backward LPC method, the buffering delay is only 5 samples, i. e. the excitation frame length. The direct extension of the LD-CELP scheme to the 20 Hz - 15 kHz bandwidth gives rise to degradations, especially coding noise and some defaults due to filter instabilities. A detailed analysis of these problems shows that the role of some coder components has to be reconsidered. The necessity for the re-optimisation of the coder parameters was confirmed by informal listening tests.

3. SYNTHESIS FILTER ORDER SELECTION

An important parameter of the system is the synthesis filter order. In the LD-CELP coding system, its value was set to 50 so that the spectral envelope and also the harmonic structure of high pitched signals could be modelled. To have the same prediction gain for a sampling rate of 32 kHz, the filter order should be equal to 200. In addition to the computation complexity caused by such a high order, we must solve the problem of the filter instability. In order to determine the optimal filter order, we measure the prediction gain depending on the filter order. Figure 1 shows an example of the experiment results. Generally, for male voices saturation occurs for a filter order greater than 60, although for the female voices this behaviour is less marked.

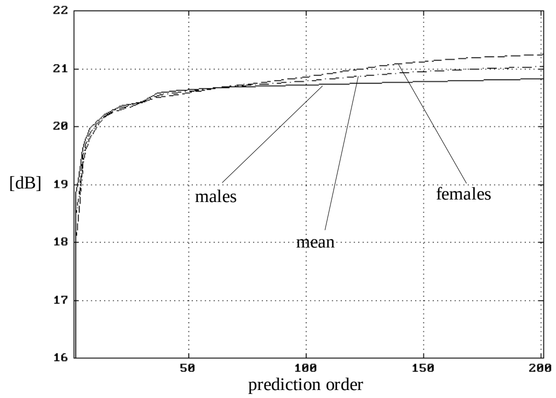

Figure 1: Behaviour of the prediction gain versus the synthesis filter order for a corpus of six male and female voices.

According to these results, we carried out a series of simulations of the coder using four different synthesis filter orders. Figure 2 illustrates the evolution of the segmental signal-to-noise-ratio (SNRseg) depending on the filter order. At an order of greater than 64 the SNRseg continues to increase slowly for music signals, but we observe a slight decrease for speech.

Thus the synthesis filter order is fixed at 64, which appears to be a fair compromise between a good spectral resolution and a reasonable algorithm complexity.

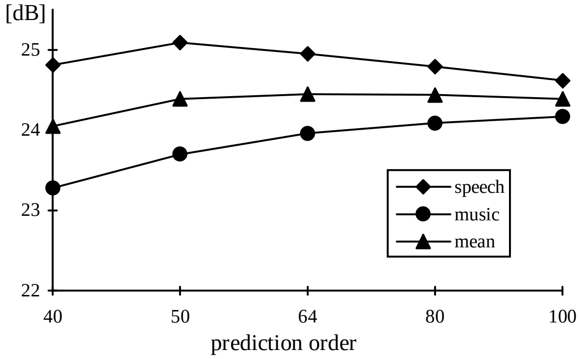

Figure 2: Behaviour of segmental signal to noise ratio depending on synthesis filter order for a corpus of ten speech and music signals.

4. RESTITUTION OF SIGNAL PERIODICITY

In our coder, the size of the excitation vector is 5 samples, which corresponds to a temporal resolution of 0.16 ms for a 32 kHz sampling frequency. This choice allows, by adapting the gain and selecting the optimum excitation vectors, the reproduction of the details and the periodicity of the linear predictive residual, without requiring a too high filter order. Therefore, all the fundamental speech frequencies can be correctly modelled. This phenomenon is illustrated in figure 3 for a female voice signal. We notice that the short term excitation presents periodic peaks fairly synchronised with the original signal pitch.

In our coder, the modelisation of the spectral envelope is provided by the synthesis filter, whereas the pitch modelisation is mainly ensured by the choice of the optimum vector and their gains.

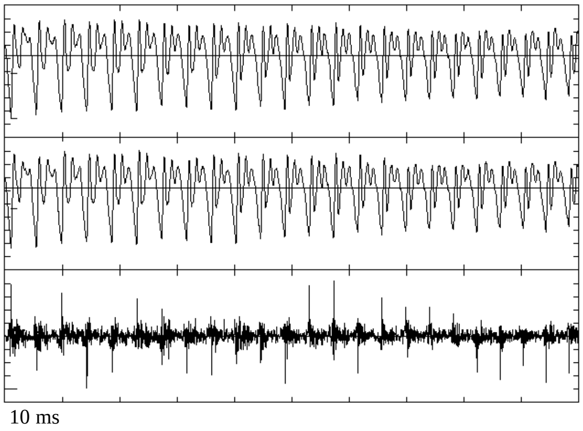

Figure 3: Modelisation of the filter synthesis excitation. From top to bottom : the original signal ($F_0 \approx 280$ Hz), the synthesised signal, the excitation of synthesis filter (order 64).

5. OPTIMUM PERCEPTUAL FILTER FOR NOISE SHAPING

In CELP coders, the use of a perceptual filter allows the shaping the coding noise spectrum, so that it could be perceptually masked by the signal. This filter has the form:

$$W(z) = A(z/\gamma_1)/A(z/\gamma_2) \qquad (1)$$

where 1/A(z) is the synthesis filter. By varying γ_1 and γ_2, it is not possible to control both the shape and the tilt of the perceptual filter spectrum on a wide bandwidth with a large dynamic range. For this reason, an improvement of the spectral modelisation has been suggested: W(z) is used in cascade with another filter controlling the spectrum tilt [3, 4]. In addition, this form of filter is only partially able to take into account the characteristics of human hearing system since it comes from an AR model of the speech production process.

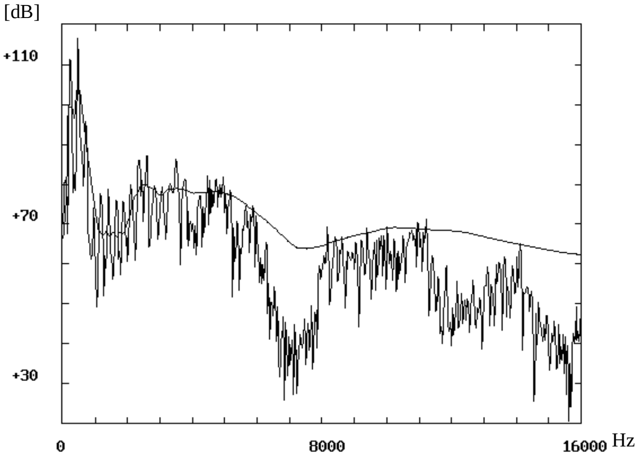

Figure 4: Speech signal spectrum and its optimum masking shape.

To solve this noise shaping problem for the 20 Hz - 15 Hz bandwidth, we propose a new optimal method for noise shaping based on a psycho-acoustic model [6, 7]. In order to avoid introducing any additional delay, we determine the perceptual filter from the previous input signal samples. At the beginning of a new LPC frame (every 2.5 ms), a masking shape is calculated on the last 1024 samples. This block length is necessary to guarantee a good spectral resolution. The samples are weighted by a hybrid analysis window, in which the more recent are weighted by a cosine window and the others by an exponential window. This increases the importance of the most recent samples in determining the filter coefficients. Furthermore, its spectral characteristics are comparable to those of the Hamming window [5]. The power spectrum of this weighted signal is calculated using the FFT algorithm. Then the optimum masking shape is obtained by the convolution (in the Bark domain) of this spectrum with the basilar membrane spreading function [6, 7]. The autocorrelation coefficients of the masking shape are obtained by an inverse FFT. Finally, the AR model is determined using the Levinson-Durbin algorithm.

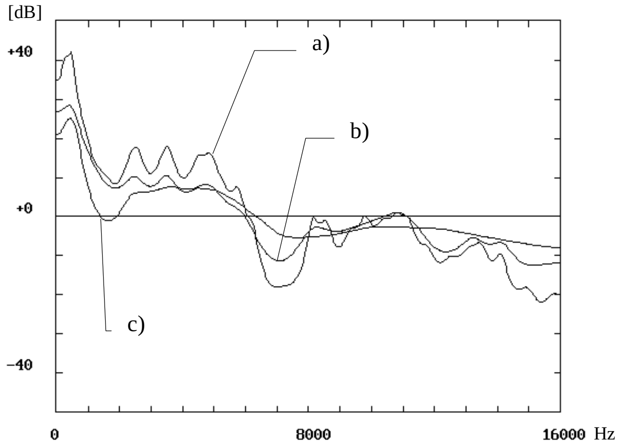

Figure 5: Behaviour of LPC spectra: a) synthesis filter; b) perceptual weighting filter determined in the classical way with tilt correction; c) perceptual filter determined from the optimum masking shape.

It should be noted that the proposed method does not introduce any additional delay, since the filter coefficients are determined from the input signal preceding the current frame.

Figure 4 shows a speech signal spectrum and its optimum masking shape obtained using the algorithm described in [6] and [7]. Figure 5 shows the LPC spectrum of the synthesis filter, the spectrum of the classical noise-weighting filter, with tilt correction, and finally the spectrum of the perceptual filter, determined from the optimum masking shape. We observe that the new LPC perceptual filter is a good model of the optimum masking shape. We also notice that the classical weighting filter spectrum is very far from the optimum coding noise shape, in the perceptual sense. In fact, between 1 kHz and 3 kHz the frequential response of the classical weighting filter is above the optimum perceptual shape, which makes the quantization noise more audible. Moreover, between 6 kHz and 8 kHz, and beyond 11 kHz, its frequency response is well below the auditory threshold. A re-optimisation of classical filter parameters (γ_1, γ_2) and tilt control filter parameters does not achieve optimum noise shaping.

6. BIT ALLOCATION AND CODEBOOK OPTIMISATION

This procedure is fully backward adaptive, so only the excitation indexes are transmitted to the decoder. The shape codebook index is coded with 7 bits and the gain codebook index with 3 bits. With a 5-sample frame and at the sampling frequency of 32 kHz the bit rate is 64 kbit/s.

The shape codebook has been designed using a closed-loop iterative procedure [8]. Figure 6 shows the results of the codebook design. Even if the W-SNRseg gain is only 0.5 dB - the starting codebook has already high performances - the increase in perceived quality is very significant. Informal listening confirms these results. The original signal is first presented to the listeners and followed - in a random order - by two synthesised signals coded with or without the optimised codebook. The listeners found the signal coded with the newly designed codebook as the nearest to the original one in 70% of cases.

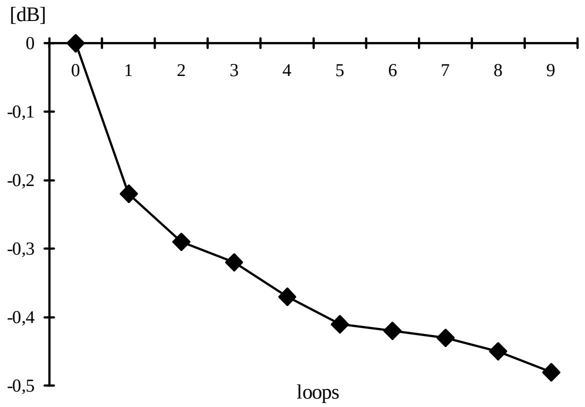

Figure 6: Results of the shape codebook optimisation design: behaviour of average weighted mean-square error depending on the optimisation loops.

7. CODER COMPLEXITY REDUCTION

In CELP coders the most computationally complex block is the analysis-by-synthesis research of the optimum code word in the codebook. The index of the optimum code vector is obtained by minimising the following distortion expression:

$$D(c_k) = \|Hc_k\|^2 - 2\langle t, Hc_k\rangle \qquad (2)$$

where c_k is the vector of index k, t the target vector, H the lower triangular matrix of the truncated impulse response of the cascade of the synthesis and perceptual filters and the symbol < , > represent the inner product. In this research procedure, for every LPC frame, the codebook vectors are filtered and in this way adjusted depending on the signal characteristics. This filtering operation implies an important computational load in the research procedure. The introduction of a ternary algebraic codebook reduces the search computational complexity [9]. This algebraic code is generated using modulo-3 arithmetic with five digits and by using the following transposition: {0, 1, 2} ⇔ {-1, 0, +1}. In this way the code-vectors with index $k = 3^i$, only have a non-zero pulse of -1 in the position i. If $S(k)$ is the last component of the vector Hc_k, it can be demonstrated that:

$$\begin{cases} S(3^i) = -h(i) \\ S(3^i - j) = -h(i) - S(j) \\ S(3^i + j) = -h(i) + S(j) \end{cases} \qquad (3)$$

where $1 \le j \le (3^i-1)/2$ and $0 \le i < 5$. We observe that the code-vectors are obtained by shifting to the left and by a link with one of the ternary symbols. Thus, the energy $E_k = ||Hc_k||^2$ can be calculated using the expression:

$$E_k = E_{\mathrm{int}(k/3+0.5)} + S^2(k) \qquad (4)$$

where int(x) is the smallest integer that is less than or equal to x and for $1 \le k \le (3^5-1)/2$. To complete the computation of the distortion the following equation can be applied:

$$\begin{cases} corr\left(3^i\right) = -v(i) \\ corr\left(3^i - j\right) = -v(i) - corr\left(j\right) \\ corr\left(3^i + j\right) = -v(i) + corr\left(j\right) \end{cases} \tag{5}$$

where $1 \le j \le (3^i-1)/2$ and $0 \le i < 5$ and $corr(k) = <t, Hc_k> = <H^T t, c_k>$ and $v=H^T t$. The computational gain obtained applying this algorithm is 28% of the total coding time.

Table I summarises the SNRseg measure on a corpus of speech and music signals when we use the optimised compared to the algebraic codebook in the analysis-by-synthesis procedure. We notice that the SNRseg is slightly degraded in the case of the ternary algebraic codebook, but informal listening tests show that the perceptual difference is insignificant.

[dB]	Speech	Music	Total
Optimised	**23.41**	**24.34**	**23.79**
Algebraic	**22.90**	**23.96**	**23.33**
Gain	**-0.51**	**-0.38**	**-0.46**

Table I: Segmental SNR obtained on a corpus of speech and music signals using two optimised versions of the coder with: optimised and algebraic codebook.

8. RESULTS

Two versions of the coder have been tested and all the parameters (synthesis filter order, excitation vector size, shape codebook) have been optimised. In the first version, we use the classical weighting filter, and in the second, the perceptual filter determined from the optimum masking shape. The segmental SNR values for a corpus of speech and music signals, synthesised by the two coders are shown in table II. By using the new perceptual filter, the segmental SNR increases by about 1.5 dB. This gain appears to come from a better noise shaping, especially in low frequencies, as illustrated in figure 4 and figure 5. Our informal listening tests confirm the quality improvement due to this new optimum-shaping method. In particular, a significant increase in the clearness of the synthesised signals is obtained.

[dB]	Speech	Music	Total
A)	**23.41**	**24.34**	**23.78**
B)	**24.62**	**26.18**	**25.24**
Gain	**+1.21**	**+1.84**	**+1.46**

Table II: Segmental SNR obtained using two optimised versions of the coder with: A) classical weighting filter and tilt correction filter; B) perceptual filter using the optimum masking shape.

CONCLUSION

We have proposed a new 64 kbit/s coding algorithm of 20 Hz - 15 kHz speech, with a very low delay of 0.6 ms. The quality obtained is quasi-transparent. We have used a new optimum-noise shaping filter based on a psycho-acoustic model of human perception system. This new filter improve significantly the speech quality. Informal listening tests on a large speech corpus and some music signals show that the coding quality is close to transparency. Finally, a modified algorithm has been proposed, allowing an important reduction of coder computational complexity, without decreasing the perceived quality of signals.

CD-ROM SOUND LINKS

Original female speech signal: [SOUND A217S1.WAV].
Coded female speech signal: [SOUND A217S2.WAV].
Original male speech signal: [SOUND A217S3.WAV].
Coded male speech signal: [SOUND A217S4.WAV].

REFERENCES

1. J.-H. Chen, R. V. Cox, Y.-C. Lin, N. Jayant & M. J. Melchner, " A Low-Delay CELP Coder for the CCITT 16 kb/s Speech Coding Standard " - IEEE Jour. on Selec. Areas in Comm., vol. 10, n° 5, pp. 830-849, 1992.

2. C. Murgia, G. Feng, C. Quinquis & A. Le Guyader, " Very Low Delay an High Quality Coding of 20 Hz - 15 kHz speech at 64 kbit/s" - Proc. of EUROSPEECH ‘95, pp. 37-40.

3. E. Ordentlich & Y. Shoham, " Low-Delay Code-Excited Linear-Predictive Coding of Wideband Speech at 32 kbps " - Proc. of ICASSP ‘91, pp. 9-12, 1991.

4. O. Gottesman & Y. Shoham, " Real -Time Implementation of High-Quality 32 kbps Wideband LD-CELP Coder " - Proc. of EUROSPEECH ‘93, pp. 1115-1118, 1993.

5. J.-H. Chen, Y.-C. Lin & R. V. Cox, " A Fixed - Point 16 kb/s LD-CELP Algorihm " - Proc. of ICASSP ‘91, pp. 21-24, 1991.

6. Y. Mahieux & J. P. Petit, " Transform Coding of Audio Signals at 64 kbit/s " - Globecom ‘90, pp. 518-522, 1990.

7. Y. Mahieux & J. P. Petit, " High-Quality Audio Transform Coding at 64 kbps " - IEEE Trans. on Comm., vol. 42, n° 11, pp. 3010-3019, 1994.

8. G. Davidson, M. Yong & A. Gersho, " Real-Time Vector Excitation Coding of Speech at 4800 bps " - Proc. of ICASSP ‘87, pp. 2189-2192, 1987.

9. R. Di Francesco, " Codage Algébrique de la parole : prédiction linéaire à excitation par code ternaire " - Ann. Télécommun. 47, n° 5-6, pp. 214-226, 1992.