Strategies to improve the performance of very low bit rate speech coders and application to a variable rate 1.2 kb/s codec

R.C. de Lamare and A. Alcaim

Abstract: This paper presents several strategies to improve the performance of very low bit rate speech coders and describes a speech codec that incorporates these strategies and operates at an average bit rate of 1.2 kb/s. The encoding algorithm is based on several improvements in a mixed multiband excitation (MMBE) linear predictive coding (LPC) structure. A switched-predictive vector quantiser technique that outperforms previously reported schemes is adopted to encode the LSF parameters. Spectral and sound specific low rate models are used in order to achieve high quality speech at low rates. An MMBE approach with three sub-bands is employed to encode voiced frames, while fricatives and stops modelling and synthesis techniques are used for unvoiced frames. This strategy is shown to provide good quality synthesised speech, at a bit rate of only 0.4 kb/s for unvoiced frames. To reduce coding noise and improve decoded speech, spectral envelope restoration combined with noise reduction (SERNR) postfilter is used. The contributions of the techniques described in this paper are separately assessed and then combined in the design of a low bit rate codec that is evaluated against the North American Mixed Excitation Linear Prediction (MELP) coder. The performance assessment is carried out in terms of the spectral distortion of LSF quantisation, mean opinion score (MOS), A/B comparison tests and the ITU-T P.862 perceptual evaluation of speech quality (PESQ) standard. Assessment results show that the improved methods for LSF quantisation, sound specific modelling and synthesis and the new postfiltering approach can significantly outperform previously reported techniques. Further results also indicate that a system combining the proposed improvements and operating at 1.2 kb/s, is comparable (slightly outperforming) a MELP coder operating at 2.4 kb/s. For tandem connection situations, the proposed system is clearly superior to the MELP coder.

1 Introduction

With the advent of digital cellular telephones, telephony with privacy for military purposes, and applications such as voice over IP networks (VOIP), very low bit rate speech coding algorithms have assumed an increased importance. Very low bit rate speech coders, such as mixed multiband excitation (MMBE) [1, 2] and mixed excitation linear prediction (MELP) [3], are usually based on linear predictive coding (LPC), where an excitation signal is applied to an all-pole filter representing the spectral envelope information of speech. CELP coders [4], which have become successful and popular as international standards, usually show some limitations at very low bit rates.

Most modern very low bit rate speech coding algorithms can deliver quite good speech quality at rates around 2.4 kb/s. Nevertheless, those that follow the classical vocoder principle of Atal and Hanauer [5] usually result in synthetic speech quality due to an impairment generally termed ‘buzziness’. This work presents several techniques to improve the performance of very low bit rate speech coders and details a speech codec that operates at an average bit rate of 1.2 kb/s. The encoding algorithm is based on several improvements in a mixed multiband excitation (MMBE) linear predictive coding (LPC) structure, even though these techniques can be adopted for other classes of coders, such as the one based on the CELP structure [4]. In this paper, the 1.2 kb/s variable rate speech codec recently proposed in [6] is described and examined in detail and a more complete and rigorous analysis in terms of speech quality assessment is presented. Furthermore, we describe the individual improvements provided by sound-specific modelling and synthesis techniques, LSF quantisation and spectral envelope restoration combined with noise reduction (SERNR) postfiltering. We also separately evaluate their impact on speech quality through several speech quality tests.

In the encoding structure described in this paper and briefly reported in [6], we employ a mixed multiband excitation (MMBE) [1, 2] approach to address the problem of ‘buzziness’ found in voiced frames, through splitting the speech into several frequency bands. To encode unvoiced frames, most compression algorithms reported in the literature employ noise excitation. However, this approach is not adequate to encode some non-stationary sounds such as unvoiced fricatives and stops. For this reason, we use a modelling and synthesis technique [7–9] that improves the encoding of these sounds, while it encodes unvoiced frames at only 0.4 kb/s. In this paper, we detail this modelling and synthesis technique and discuss the benefits and drawbacks of the proposed method.

IEE Proceedings online no. 20051189

doi: 10.1049/ip-vis:20051189

Paper first received 23rd March and in revised form 23rd September 2004

R.C. de Lamare and A. Alcaim are with CETUC-PUC/RIO, Rua Marquês de S. Vicente, 225-Gávea, 22453-900 Rio de Janeiro-RJ, Brazil

E-mail: delamare@infolink.com.br

Another fundamental issue for low bit rate speech coders is the pitch detection algorithm. Pitch detectors are responsible for the computation of the pitch period as well as for the classification of voiced and unvoices frames. In the encoding scheme described in this paper, we have chosen a strategy based on the pitch detection algorithm reported in [7], that uses a sliding window to further reduce incorrect pitch values and voicing decisions. We also employ a classification algorithm that distinguishes voiced, unvoiced fricatives, unvoiced stops and silence frames.

To represent the LPC coefficients we have chosen the line spectral frequencies (LSF) [10]. In memoryless vector quantisation (MVQ) [11–14], each LSF vector is quantised independently of any other LSF set. However, this is not the most efficient approach to encode LSF parameters, since large gains can be achieved by exploiting the inherent interframe correlation between adjacent LSF vectors, especially for voiced segments. A number of predictive vector quantisation (PVQ) [13, 14] and switched-predictive vector quantisation (SPVQ) [15–17] schemes, which exploit interframe correlation, have been proposed in the last few years. In this work, we employ an enhanced SPVQ scheme [18, 19] that outperforms previously reported structures, to encode the LSF parameters. A procedure to jointly optimise the codebooks [20] is then used to improve the performance of the LSF quantiser employed in the proposed speech codec.

The adaptive spectral enhancement (ASE) filter, proposed by Chen and Gersho [21], is one of the most popular and successful postfiltering techniques. This strategy reduces the spectral components of the decoded speech signal that exhibit low signal-to-noise ratio. Another strategy to enhance the quality of decoded speech, the spectral envelope restoration (SER) filter, introduced by da Silva and Alcaim [22], attempts to reconstruct the short-time spectral envelope (*stse*) of speech. The principle of this postfilter is to remove from the reconstructed speech its *stse* and apply the *stse* obtained from the received LPC parameters. The SER approach has been shown to reproduce speech with a quality comparable to the ASE technique. Here, we describe a spectral envelope reconstruction combined with noise reduction (SERNR) postfilter [9], that combines the strengths of the ASE and the SER strategies. The SERNR postfilter has the spectral envelope restoration properties of the SER filter and the noise reduction capabilities of the ASE technique. The SERNR can significantly enhance decoded speech quality and has shown a performance superior to the traditional ASE filter for MMBE platforms [9, 23]. To reduce background noise and enhance the quality of the synthesised speech, we employ a noise suppression method, proposed by Arslan *et al.* [24], that is based on a smooth spectral subtraction of noise-corrupted speech.

The performance of the sound-specific modelling and synthesis, LSF quantisation and SERNR postfiltering are individually assessed and the 1.2 kb/s speech coder is compared to the North American standard MELP coder [3], in terms of both objective and subjective listening tests. The codecs are also evaluated in tandem connection situations, where the speech signals are encoded and decoded more than one time. We present objective tests in terms of spectral distortions and percentage of outliers of the LSF quantisation and the recently adopted ITU-T P.862 perceptual evaluation of speech quality (PESQ) recommendation. Subjective listening tests results with mean opinion scores (MOS) and A/B comparison tests are also reported.

2 Excitation model

In this Section the excitation model employed in the experiments throughout this work and its components are described. Section 2.1 describes the sliding window pitch detection scheme, Section 2.2 details our mixed-multiband excitation approach; and Section 2.3 presents the fricatives and stops modelling and synthesis strategies.

2.1 Pitch detection

Pitch detection is one of the most important issues in low bit rate coding, since it has a significant impact upon the quality of synthesised speech. Indeed, the more accurately a pitch algorithm detects the pitch period and decides voicing, the more it contributes to the quality of reproduced speech. Pitch estimation algorithms are responsible for computation of the pitch period and classification of voiced and unvoiced frames. For this reason, we have chosen a strategy based upon the pitch detection algorithm introduced by Unno *et al.* [7], which uses a sliding window to further reduce incorrect pitch values and voicing decisions. The deployment of a sliding window can reduce the artificial noise usually found in non-stationary segments that contain vowels and result in more accurate voicing decisions and pitch estimates. The pitch correlation provided by the sliding window method is defined by

$$R(T) = max_{i=-T_s}^{T_s-1}[max_T R_i(T)] \tag{1}$$

$$R_i(T) = \frac{C(i, T+i)}{\sqrt{C(i,i)C(T+i, T+i)}} \tag{2}$$

where T_s is the maximum sliding range and $R_i(T)$ is the value of the normalised autocorrelation for the delay i. The autocorrelation function $C(k,l)$ is limited between 20 and 160 samples (at a sampling frequency of 8 kHz) and is expressed by

$$C(k,l) = \sum_{n=0}^{N-1} s(n+k)s(n+l) \tag{3}$$

where $s(n)$ is the lowpass speech signal, N is the frame size and k and l are the corresponding delays.

2.2 Mixed multiband excitation

Mixed multiband excitation (MMBE) [1, 2] addresses the problem of ‘buzziness’ found in low bit rate coders that follow the classical vocoder principle, through splitting the speech into several frequency bands. These frequency bands have their voicing assessed individually, with a voiced excitation source or an unvoiced excitation source for each subband in the speech frame. This excitation model is capable of representing the voiced frames more adequately than those models which assume a binary decision between voiced and unvoiced frames [5]. To apply this model at very low bit rates it is necessary to establish a trade-off between speech quality and the number of subbands used in the coder [1–3]. Taking into account this trade-off, and in order to achieve an average bit rate of 1.2 kb/s, we have chosen a model with only three subbands, resulting in four mixed excitations to represent voiced signals.

The subband analysis filters split the speech spectrum into the following frequency bands: 0–1 kHz, 1–2 kHz and 2–4 kHz. Figure 1 shows the frequency response of these filters. We have employed the same method used in the MELP coder [3] to select voiced excitation. The bandpass voicing analysis uses the pitch detector to determine

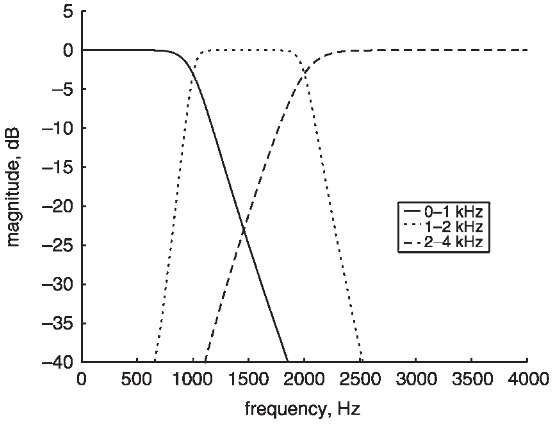


Fig. 1 *Frequency response of the analysis filter bank*

whether each one of the analysed subbands is voiced or unvoiced. This approach is undertaken for those frames which are declared voiced by the pitch detector. The procedure is to filter the input speech signal into the frequency bands 1–2 kHz and 2–4 kHz and perform pitch detection on each of the subbands. The voicing decision $v_i \in \{0, 1\}$, where the numbers 0 and 1 denote unvoiced detection, respectively, and i denotes the ith sub-band with $1 \leq i \leq 3$, is the largest value of the pitch correlation $R(T)$ defined in (1), computed for both the bandpass signal and the time envelope of the bandpass signal. The time envelope is first decremented by 0.1 to compensate for an experimentally observed bias as detailed in [3]. The envelopes are calculated by full-wave rectification followed by a smoothing filter. This filter consists of a zero at DC in cascade with a complex pole pair at 150 Hz with a radius of 0.97. If the frame is found to be voiced by the pitch detector, the voicing decision of the first band is automatically set to $v_1 = 1$ and the remaining ones, for $i = 2, 3$ are given by

$$v_i = \begin{cases} 1 & \text{if } R(T) > 0.6 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

The proposed excitation codebook, shown in Table 1, has four different mixed-voiced excitations (0–3) and four unvoiced excitation (4–7) entries in the codebook, which are used to encode fricatives, stops and silence frames, totalling 3 bits for the excitation. The mixed excitation $m(n)$ used in this work consists in the application of a pulse excitation $p(n)$ filtered in the chosen subbands and summed to filtered white noise excitation $w(n)$ in the remaining subbands. Indeed, the mixed excitation $m(n)$ is computed by summing the filtered pulse and noise excitations, i.e. $p(n)$ and $w(n)$, as given by

$$m(n) = p(n) + w(n) \tag{5}$$

where $p(n)$ and $w(n)$ are obtained via the following filtering operations and according to the excitation index:

$$p(n) = \begin{cases} t(n) - \sum_{k=1}^{N_{0-4}} h_k^{0-4} p(n-k) & \text{if index} = 0 \\ t(n) - \sum_{k=1}^{N_{0-2}} h_k^{0-2} p(n-k) & \text{if index} = 1 \\ t(n) - \sum_{k=1}^{N_{0-1}} h_k^{0-1} p(n-k) & \\ \quad - \sum_{k=1}^{N_{2-4}} h_k^{2-4} p(n-k) & \text{if index} = 2 \\ t(n) - \sum_{k=1}^{N_{0-1}} h_k^{0-1} p(n-k) & \text{if index} = 3 \end{cases} \tag{6}$$

$$w(n) = \begin{cases} 0 & \text{if index} = 0 \\ u(n) - \sum_{k=1}^{N_{2-4}} h_k^{2-4} w(n-k) & \text{if index} = 1 \\ u(n) - \sum_{k=1}^{N_{1-2}} h_k^{1-2} w(n-k) & \text{if index} = 2 \\ u(n) - \sum_{k=1}^{N_{1-4}} h_k^{1-4} w(n-k) & \text{if index} = 3 \end{cases} \tag{7}$$

where $t(n)$ is a pulse train whose pulses are spaced by the pitch period, $u(n)$ is a white Gaussian noise with zero mean and unit variance, and h_k^B is the kth coefficient for the Bth band of the synthesis filter bank. The filter bank contains infinite impulse response (IIR) filters of the ellyptical type with 0.5 dB of ripple in the passband, at least 60 dB of attenuation in the stopband and N_Bth-order filters. For bandpass filters, we have used $N_B = 12$, whereas for lowpass and highpass filters $N_B = 6$.

At the decoder, the excitation is generated with the aid of a pair of filter banks as given by (1)–(4), which are different from those used in the analysis. Note that at the encoder we only wish to determine the voiced subbands, determined on the basis of the pitch detection algorithm invoked for each subband signal. Differently from the MELP that employs 64th-order finite impulse response (FIR) filters, in the proposed coder, the filters are IIR and were designed in order to minimise the number of filters used in the process of synthesis of the mixed excitation signal. For example, if we transmit index 1 for a given frame, the MELP employs three FIR filters (out of the five connected in parallel) in the 0–0.5, 0.5–1 and 1–2 kHz subbands to filter the periodic excitation $p(n)$, whereas the proposed coder uses only one IIR filter in the 0–2 kHz subband for the synthesis. Note that in such situations, where we have adjacent voiced excitations, the MELP employs unnecessary filtering operations that may introduce undesirable distortions. Indeed, informal listening tests have shown that in the cases where we transmit index 1 for a given frame, our approach provides a higher quality speech synthesis than that attained using the MELP synthesis filters. Figure 2 shows the frequency response of the synthesis filters for periodic and for noise excitation.

2.3 Fricatives and stops encoding

The noise excitation usually employed to model unvoiced sounds is not capable of adequately representing unvoiced stops and fricatives. In order to provide a clearer speech quality for the sentences containing stops and fricatives sounds, we use a strategy based upon the algorithms recently introduced in [9].

The methods reported in [9] resemble those introduced in [7] and [8]. However, the overall scheme has novel contributions with respect to some specific points that differ

Table 1: Proposed mixed-excitation codebook

Index	Band for $p(n)$, kHz	Band for $w(n)$, kHz	Index	Unvoiced excitation
0	0–4	–	4	stop 1
1	0–2	2–4	5	stop 2
2	0–1 and 2–4	0–4	6	fricative
3	0–1	1–4	7	silence

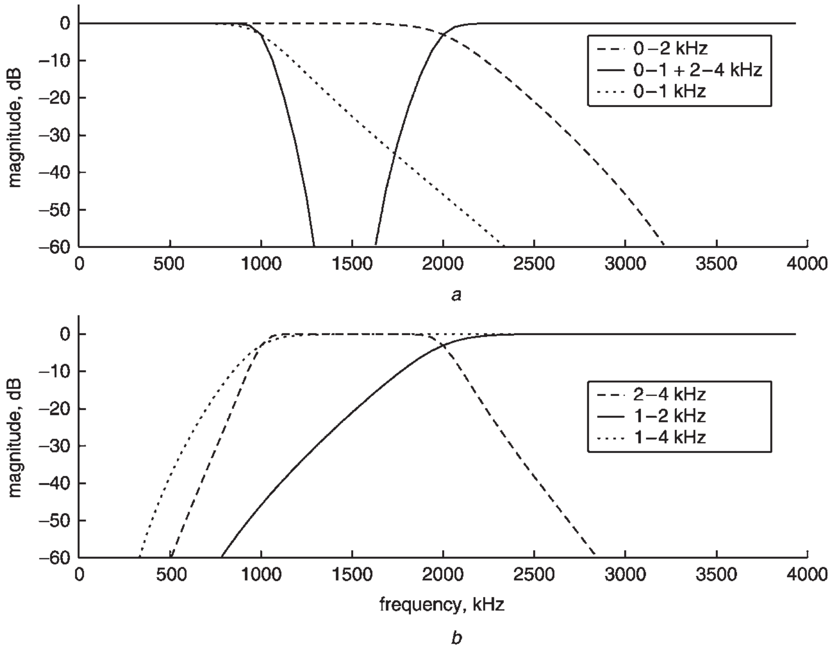


Fig. 2 *Frequency response of the synthesis filter banks*

from the techniques described in [7] and [8]. Regarding the modelling and synthesis of stops, the work by Unno *et al.* [7] employed pre-stored excitation templates and transmitted LPC parameters to model and synthesise stop sounds. In [8] Ehnert used pre-stored excitation templates of 40 ms duration and pre-stored LPC sets to reproduce stop sounds. This method requires no transmission of LPC parameters, however, according to informal listening tests it has shown inferior performance to Unno *et al.*'s. In [9] we proposed a modification of the techniques introduced by Unno *et al.* to encode stops that rather than transmitting LSF vectors, makes use of pre-stored templates of LPC coefficients. Informal listening tests have shown that our approach is capable of reproducing speech at a quality comparable to that delivered by the method in [7], while reducing the bit rate to only 0.4 kb/s and offering an attractive trade-off between speech quality and bit rate.

To model and reproduce fricative signals, Ehnert [8] used pre-stored excitation templates of 220 ms duration and pre-stored LPC sets to reproduce these sounds. For the case of fricative sounds, the difference of our work in [9] from [8] is that the pre-stored excitation and LPC set templates correspond to only 20 ms of speech. At first glance, one may think that this idea would lead to bad modelling because of the different types of fricatives. However, distinctive characteristics of these sounds will be captured in the transition of the fricative to the voiced sounds and vice versa, where a mixed excitation will be used. Specifically, when different fricative sounds are encoded the mixed excitation that precedes or follows the template model will contribute to account for their differences, even though some speech quality deterioration is assumed. It should be noted that a small percentage of unvoiced speech segments have to rely on mixed-voiced excitation to compensate for different fricatives and stops. The overall bit rate is not significantly affected by this approach as shown in Section 5 by the statistical analysis carried out to determine the average bit rate of the proposed codec. Certainly, the use of a reduced set of templates is not capable of achieving high quality speech, although they play a key role in reducing the overall bit rate of the codec. In this context, one can employ a specific template for each kind of fricative and stop sound, leading to an increase in the bit rate. In this regard, such an approach is still an open topic since it requires more sophisticated detection methods in order to discriminate different types of fricatives and stops. From informal listening tests we have verified that the subjective quality of reconstructed fricative sounds can be improved by reducing the length of these pre-stored templates. When successive fricative frames are detected, our approach uses the same pre-stored template with the appropriate gain for each one of the frames in order to reproduce the fricative sound.

For the detection of stop sounds we employ the peakiness value of the LPC residual signal $r(n)$ and a sliding window is used to find the frame position that maximises the peakiness value. The peakiness value with the sliding window is given by

$$P = max_{i=-T_s}^{i=T_s} \frac{\frac{1}{N}\sum_{n=0}^{N-1} r(n+i)^2}{\sqrt{\frac{1}{N}\sum_{n=0}^{N-1} |r(n+i)|}} \qquad (8)$$

where N is the frame size and T_s is the maximum sliding range. In this approach there are two types of stop signals since two excitation codebook entries are reserved for these sounds. The first one corresponds to those signals whose maximum amplitudes are located in the first half of the frames while the second one is associated to those whose maximum amplitudes are found in the second part.

The detection of fricative sounds makes use of appropriate thresholds for the zero crossings and the energy of each frame. These low energy signals usually have between 60 and 140 zero crossings per frame while voiced frames typically do not cross the axis more than 60 times per frame [8].

Despite an accurate assessment (including insertion and deletion errors) of the efficiency of the detection schemes not being carried out, these techniques have been shown to work well for a wide range of situations of practical interest. Indeed, we focused on the applicability of these methods to an excitation model with fricatives and stops. We believe that a more detailed study of detection of fricatives and stops is beyond the scope of this paper, even though it constitutes an interesting research topic. It is also important to remark that in the situations where these sounds are not identified our scheme has the option of choosing among four mixed-voiced excitations.

For the reproduction of unvoiced stops and fricatives, we employ a model where fricatives (*f*) and stops (*s*) signals ($f|s(n)$ is used to denote these sounds) are produced by scaling and LPC filtering pre-stored templates of LPC residual signals $r(n)$ and LPC coefficients $(a_i,\ i = 1, \ldots, p)$:

$$f|s(n) = Gr(n) + \sum_{i=1}^{p} a_i f|s(n-i) \qquad (9)$$

where G is a gain based on the energy of the input stop or fricative signal and $\{a_i\}$ is the set of LPC coefficients stored

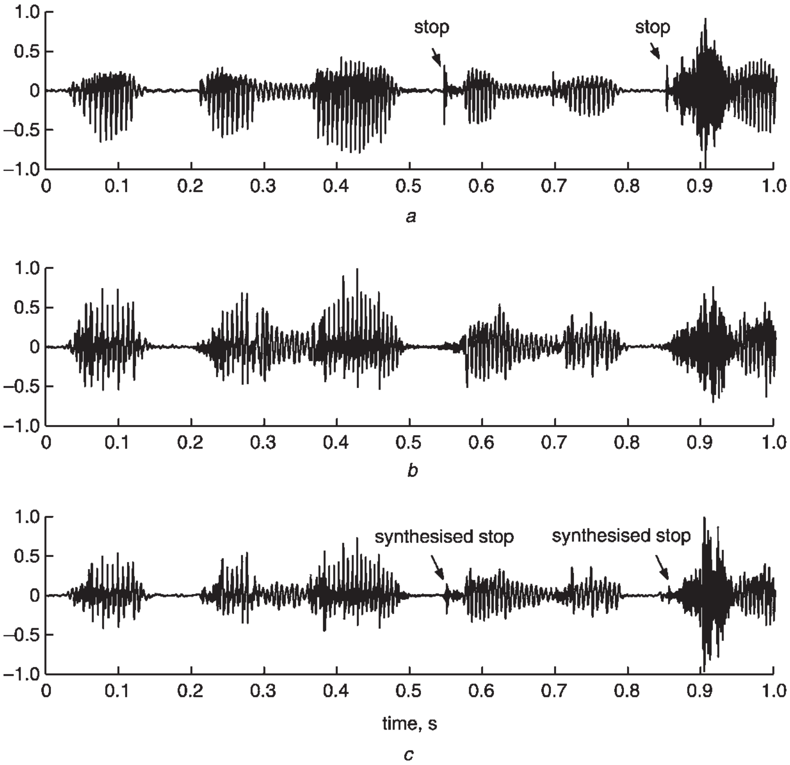


Fig. 3 *Reproduction of stop signal in synthesised speech*

a Input speech signal
b Synthesised speech signal with noise excitation
c Synthesised speech signal with stops modelling and synthesis

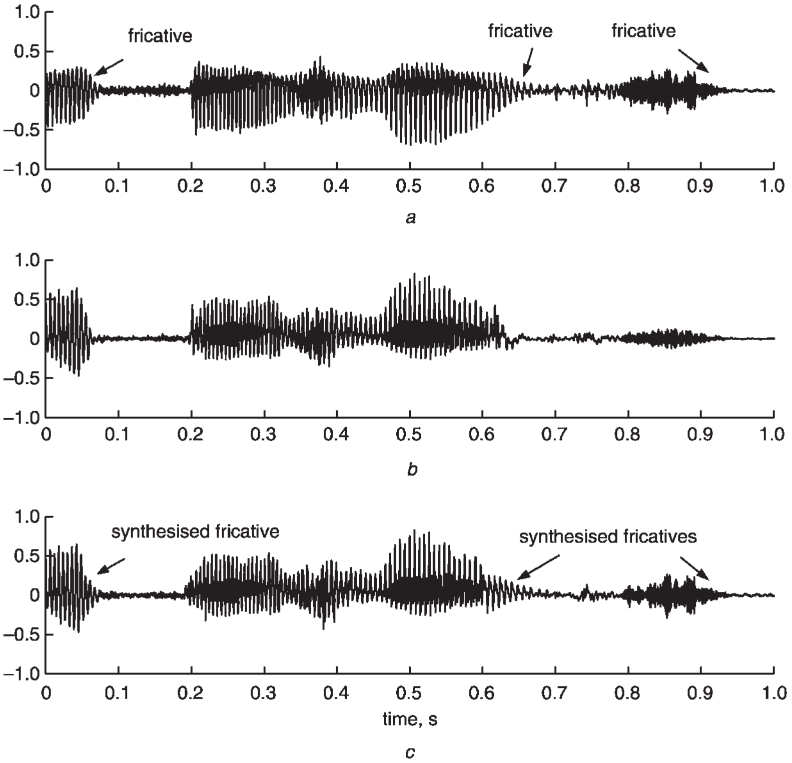


Fig. 4 *Reproduction of fricative signal in synthesised speech*

a Input speech signal
b Synthesised speech signal with noise excitation
c Synthesised speech signal with fricatives modelling and synthesis

in the decoder. The templates are carefully chosen from fricatives and stops speech segments to avoid the transmission of LPC coefficients for unvoiced frames. With regard to the selection of templates, we selected several small sized stop and fricatives residual signals for use as candidate templates. Then, we carried out several listening tests with these selected candidates in order to determine the most appropriate ones for use in the proposed model. These templates have been shown to reproduce good quality speech through informal listening tests. We used one residual signal and an LPC set as templates to synthesise fricatives, while two residual signals and two LPC sets were employed to reproduce stops. These sounds are reproduced by the application of (9), where the $r(n)$ and $\{a_i\}$ templates are used with the transmitted gains for the synthesis.

Figure 3 shows an example of speech containing stops. The input speech signal is depicted in Fig. 3*a*. In the synthesised speech signal (b) a noise excitation is applied to the full band of the speech segment associated with the stop sound. From informal listening tests it was perceived that this latter approach degrades the clarity of the speech quality. In Fig. 3*c* the synthesised speech signal provides better stops reproduction and encodes frames with only 0.4 kb/s. Figure 4 shows an example of speech containing fricatives. The input speech signal is depicted in Fig. 4*a*. The synthesised speech signal shown in Fig. 4*b* is obtained from noise excitation applied to the full band of the speech segment associated with the fricative sound. In Fig. 4*c* the synthesised speech signal provides better fricatives reproduction and encodes frames with only 0.4 kb/s. Note that the proposed scheme for the encoding of fricative and stop signals plays a significant role in the reduction of the average bit rate of the coder described in this work. We also remark that the fricatives and stops encoding scheme can be implemented with the transmission of LPC coefficients at the cost of increasing the transmission rate of the codec. Indeed, informal listening tests reveal that when the LPC parameters are transmitted the quality of fricative and stop sounds was found to be superior to the pre-stored templates approach. In addition, statistical analysis (detailed in Section 5 and carried out to determine the average bit rate of the analysed codec) indicates that unvoiced frames account for 40% of the frames. These results show the importance of the contribution of the fricatives and stops encoding scheme in terms of the average bit rate of the codec.

3 LSF quantisation

Most low bit rate speech coding algorithms are based on linear predictive coding (LPC), where an excitation signal is applied to an all-pole filter representing the spectral envelope information of speech. We have chosen line spectral frequencies (LSF) to represent LPC coefficients since they have proven to be a suitable representation of the spectral envelope and because they are well suited to quantisation and interpolation. In addition, LSF parameters usually show significant correlation between successive frames, especially for voiced segments.

Multistage vector quantisation (VQ) and split VQ are some of the most usual and successful suboptimal schemes used to encode LSF parameters. Multistage VQ has been shown to perform better than split VQ, at the expense of higher computational complexity [12, 14]. In a multistage VQ system the LSF vector $\mathbf{f}$ is approximated by the quantised vector $\hat{\mathbf{f}}$ given by

$$\hat{\mathbf{f}} = \mathbf{c}_{1i} + \mathbf{c}_{2l} + \cdots + \mathbf{c}_{Kv} \qquad (10)$$

where K is the number of stages and $\mathbf{c}_{ki}$ is the ith codevector of the codebook of the kth stage represented by the set $\mathbf{C}_k = \{\mathbf{c}_{ki}, i = 1, \ldots, I_k\}$, where I_k is the number of LSF codevectors stored in each codebook.

In memoryless vector quantisation (MVQ), each LSF vector is quantised independently of any other LSF set [11]. Paliwal and Atal [11] demonstrated that a split MVQ scheme is capable of efficiently encoding the LSF parameters with 24 bits per frame. A tree-structured multistage MVQ was presented in [12] and shown to outperform the split MVQ. However, this is not the most efficient approach to encode LSF parameters, since large gains can be achieved by exploiting the inherent interframe correlation between adjacent LSF vectors. A number of predictive vector quantisation (PVQ) schemes, which benefit from the interframe correlation, have been proposed in recent years [14–16]. A vector linear predictor forms an estimate of the incoming vectors as a linear combination of earlier observations, and the prediction residual vector (δ_{j+1}) is vector quantised. The vector δ_{j+1} is expressed by

$$\delta_{j+1} = \mathbf{f}_{j+1} - \hat{\mathbf{f}}_j \cdot \rho^t \qquad (11)$$

where ρ is the vector with the correlation coefficients and $\hat{\mathbf{f}}_j$ is the quantised version of $\mathbf{f}_j$, the LSF vector occurring at time instant j. In this work, we restrict the experiments to first-order predictors, which have been shown to capture most of the achievable coding gains.

Interframe correlation can be exploited by memory VQ methods such as PVQ. However, there are situations of rapid changes in the spectral envelope and hence low correlations between adjacent LSF vectors. Indeed, this observation motivated the combination of MVQ and PVQ techniques for encoding low correlation frames separately from typical highly correlated frames. A search of both VQ schemes is performed for each frame and the best candidate, with respect to a distortion criterion, $d(\mathbf{f}, \hat{\mathbf{f}})$, is encoded and transmitted [15].

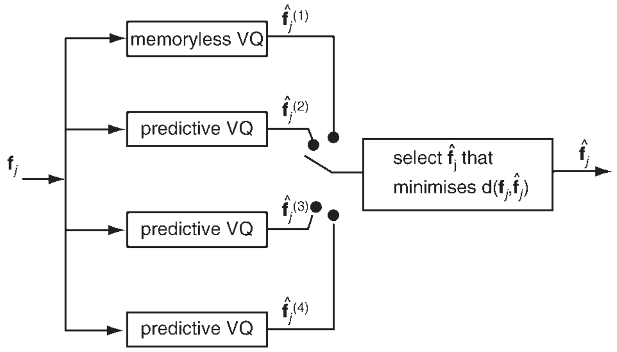
Several switched-predictive VQ (SPVQ) schemes have been reported in the literature [15–19]. All of them share the important characteristic of exploiting the changes in interframe LSF correlations, by switching between more than one PVQ (and eventually MVQ) systems. Note that different strategies to design SPVQ structures will lead to different performances.

One of these SPVQ schemes, also called safety-net VQ, was described by Eriksson *et al.* [16]. It switches between a split MVQ and a split PVQ, achieving an efficient quantisation of the LSF parameters at 20 bits per frame. Another SPVQ technique, proposed by McCree and De Martin [17], uses $N_{VQ} = 2$ tree-structured multistage PVQ schemes, where each predictor is trained for a specific training database and operates at 21 bits per frame. Recently [18, 19], we investigated the performance of this class of algorithms and introduced two SPVQ schemes that use $N_{VQ} = 4$ tree-structured multistage vector quantisers and outperform previously reported systems. It should be remarked that the best results were achieved using a larger number of reduced dimension codebooks. We now summarise the main results of our investigation and present an optimised SPVQ structure to be used in the speech codec described in this paper. The structures of the SPVQ schemes are described in Table 2 and an SPVQ structure using three PVQs and one MVQ is shown in Fig. 5.

In SPVQ systems a search among all the VQ schemes is performed and the i_{opt}th VQ, that minimises a desired distortion criterion is selected to encode a given frame according to

Table 2: SPVQ schemes and their structure with respect to the combination of PVQS and PVQ

SPVQ	Structure of quantisers
SPVQ2	1 PVQ and 1 MVQ schemes
SPVQP2	2 PVQ schemes
SPVQ4	3 PVQ and 1 MVQ schemes
SPVQP4	4 PVQ schemes



memoryless VQ
predictive VQ
predictive VQ
predictive VQ
$\hat{\mathbf{f}}_j^{(1)}$
$\hat{\mathbf{f}}_j^{(2)}$
$\hat{\mathbf{f}}_j^{(3)}$
$\hat{\mathbf{f}}_j^{(4)}$
$\mathbf{f}_j$
select $\hat{\mathbf{f}}_j$ that minimises $d(\mathbf{f}_j, \hat{\mathbf{f}}_j)$
$\hat{\mathbf{f}}_j$

Fig. 5 *SPVQ using three PVQ and one MVQ schemes*

$$i_{opt} = min\,arg\{d(\mathbf{f}, \hat{\mathbf{f}}^{(i)})\}_{i=1,\ldots,N_{VQ}} \qquad (12)$$

where $d(\mathbf{f}, \hat{\mathbf{f}})$ is the distortion criterion and N_{VQ} is the number of vector quantisers used in the system.

The speech training database used in this work consists of 2400 s of phonetically balanced sentences [25] (in the Portuguese spoken in Rio de Janeiro) uttered by 40 speakers, 20 male and 20 female. Another set of 30 s of speech uttered by six different speakers, three male and three female, was used for assessment. The speech was lowpass filtered at 3.4 kHz, highpass filtered at 120 Hz and sampled at 8 kHz. A 10th-order LPC analysis using the autocorrelation method was performed every 20 ms using a 24 ms Hamming analysis window and 15 Hz bandwidth expansion was applied to each pole of the LPC vectors. We have chosen the weighted Euclidean measure [14, 18, 19, 26] as the distortion measure, since it has been shown to improve both objective and subjective quality of compressed speech and also because it is a low complexity measure for implementation. It is given by

$$d(\mathbf{f}, \hat{\mathbf{f}}) = \sum_{i=1}^{p} \alpha_i(\mathbf{f})(f_i - \hat{f}_i)^2 \qquad (13)$$

where $\alpha(\mathbf{f}) = (\alpha_1(\mathbf{f})\alpha_2(\mathbf{f})\ldots\alpha_p(\mathbf{f}))$ is the weighting function defined as

$$\alpha_i(\mathbf{f}) = \frac{1}{f_i - f_{i-1}} + \frac{1}{f_{i+1} - f_i} \qquad (14)$$

where $i = 1, \ldots, p, f_0 = 0$ and $f_{p+1} = 0.5$.

Large savings in complexity and storage can be achieved by the use of suboptimal procedures, such as multistage VQ and tree-structured approaches, where an M-best approximation is saved at each stage of the search procedure. Indeed, we have used a tree-structured multistage VQ scheme with $M = 12$ and 4 stages.

The VQ performance was evaluated by the spectral distance (SD) expressed by

$$SD = \left[\sum_{f=0}^{4000} \frac{1}{4000}\left(10\log 10\left|\frac{S(f)}{\hat{S}(f)}\right|\right)^2\right]^{1/2} dB \qquad (15)$$

where $S(f)$ and $\hat{S}(f)$ represent the original and quantised LPC spectral envelopes. For implementation of (15) the number of the FFT points was set to 128.

In the design of vector quantisers of different coding structures, the training database was divided into subsets according to the SD between consecutive LSF vectors. The predictors used in the PVQ schemes of the different encoding systems were designed on the basis of the correlation coefficients computed from each subset of the training LSF vectors. The same number of LSF vectors for each subset was used in the design of the SPVQP2, SPVQ4 and SPVQP4, whereas for the SPVQ2 we employed a subset with a third of the training LSF vectors for the PVQ and the remaining LSFs for the MVQ.

The noise-free channel performance is shown in Fig. 6, where the average SD of the MVQ, PVQ and SPVQ methods are plotted as a function of the number of bits per frame. From this figure, it is clear that the schemes SPVQ4 and SPVQP4 achieve the best performances among the switched-predictive techniques for noise-free channels. None of these quantisers has SD greater than 4 dB. The percentage of outliers (percentage of vectors with SD) between 2 and 4 dB for the case of 21 bits per frame is approximately 2.6% for the two schemes.

Note that in noisy channels the SPVQ4, which contains one memoryless VQ, performs better than the SPVQP4. For this reason, we have chosen the SPVQ4 as the structure to be optimised.

An improved SPVQ4 structure was designed in order to provide the analysed coding scheme with an enhanced LSF encoding. We reduced the number of stages of the SPVQ4 system from four to three and the codebooks have been jointly optimised. We used an SPVQ scheme with three PVQs and one MVQ systems, depicted in Fig. 5, and designed a tree-structured multistage VQ scheme operating at 21 bits per frame with $M = 12$ and three stages. In the SPVQ structure, 2 bits are used to switch among the three PVQs and the MVQ systems, whose stages have 7, 6 and 6 bits, respectively. Moreover, the vector quantiser codebooks were jointly optimised by an approach introduced by Barnes and Frost [20], where each codebook is retrained fixing the remaining codebooks. This procedure is capable of significantly reducing the percentage of outliers, resulting

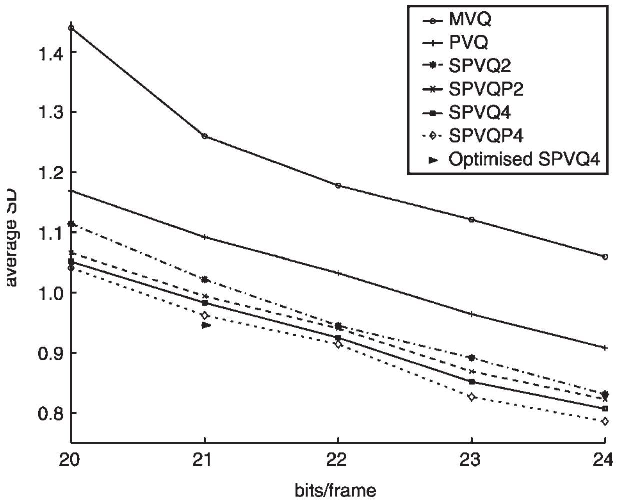


Fig. 6 *Performance of vector quantisers in terms of average spectral distance for noise-free channel*

Table 3: Performance of the 21-bit SPVQ system with jointly optimised codebooks and $M = 12$ candidates for the tree-structured search

$\bar{S}D(dB)$	0.95
$\%2-4\,\mathrm{dB}$	1.42
$\%>4\,\mathrm{dB}$	0

in superior VQ performance. The performance of the optimised SPVQ system employed in the proposed coder is shown in Table 3. For the sake of comparison, we have also shown the SD performance of this 21-bit quantiser in Fig. 6.

4 SERNR postfiltering and noise suppression

In this section we detail the postfiltering and noise suppression strategies used in our experiments. Section 4.1 describes the SERNR postfiltering technique and Section 4.2 is devoted to the noise suppression scheme.

4.1 SERNR postfiltering

One strategy to reduce the perceived coding noise is to employ an adaptive postfilter at the output of the decoder. The postfilter is designed and adapted to reduce the spectral components of the decoded speech signal that exhibit low signal-to-noise ratio. The adaptive spectral enhancement (ASE) postfilter [21] is the most usual structure and has the following transfer function:

$$H_{ASE} = \frac{A(z/\alpha)}{A(z/\beta)}(1 - vz^{-1}) \tag{16}$$

where $A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i}$ is the inverse of the synthesis filter and $\{a_i\}$ is the set of LPC parameters. Appropriate values for α, β and v at low bit rates are 0.5, 0.8 and $0.4k_1$, respectively, where k_1 is the first reflection coefficient of the linear prediction model. In a number of speech coding algorithms, this filter is followed by a fixed pulse dispersion filter (PD) that reduces some of the harsh quality of the synthetic speech.

Another strategy to enhance the quality of decoded speech attempts to reconstruct the short-time spectral envelope (*stse*) of the speech. The principle of this postfilter is to remove from the reconstructed speech its *stse* and apply the *stse* obtained from the received LPC parameters. This adaptive postfilter is called spectral envelope restoration (SER) [22] and has the following transfer function:

$$H_{SER} = \frac{\tilde{A}(z/\xi)}{A(z/\xi)} \tag{17}$$

where $\tilde{A}(z)$ is the reconstructed *stse*, obtained from an LP analysis based on the autocorrelation method, and performed on the decoded speech using a 24 ms Hamming window. $A(z)$ is the decoded *stse* an ξ must be less than 1 in order to smooth the amplitude spectrum of the postfilter. From informal listening tests we verified that the SER postfilter reproduced speech at a quality comparable to the ASE postfilter, while operating in low bit rate codecs. This fact motivated the investigation of a strategy to enhance the quality of the decoded speech which combines the strengths of the ASE and SER filters. The SERNR postfiltering structure gathers the *stse* restoration properties of the SER filter and the noise reduction capabilities of the ASE technique. The SERNR postfilter has the following transfer function:

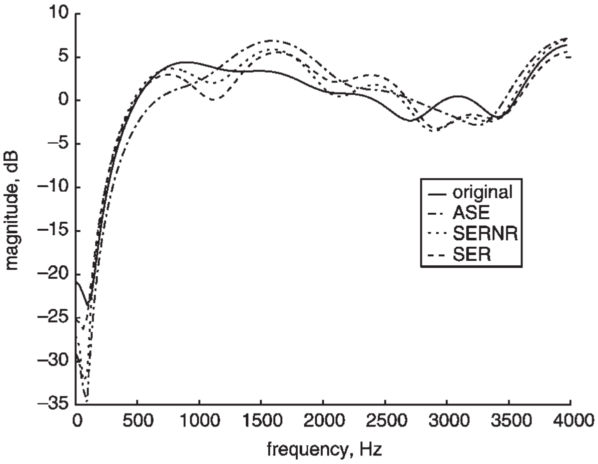


Fig. 7 *Frequency response of three different postfiltering techniques*

$$H_{SERNR} = \frac{\tilde{A}(z/\zeta)}{A(z/\eta)}(1 - vz^{-1}) \tag{18}$$

where $A(z)$ and $\tilde{A}(z)$ model the *stse* of the original and reconstructed speech, respectively. Listening tests have shown that appropriate values for ζ, η and v are 0.82, 0.9 and $0.3k_1$, respectively, and that the SERNR postfilter is superior to the SER and to the ASE postfilters. An important aspect of the SERNR and the SER postfilters performance is that they are closely related to the LSF quantiser performance because they attempt to reconstruct the *stse* obtained from the received LSFs. Therefore, it is paramount that the encoding process can deliver high quality LPC parameters in order to provide an accurate *stse* restoration and this is the case when the optimised LSF coding structure described in this paper is used.

In Fig. 7 the frequency response of a speech segment is shown for the original speech and the decoded speech using the ASE filter followed by the PD filter, the SER postfilter and the SERNR postfilter. Evaluation of these postfilters in terms of the spectral distance (SD) over several speech frames was also carried out. In particular, the SD was measured between the original LPC sets and the decoded LPC parameters after processing of the analysed postfilters and the results are depicted in Fig. 8. Note that the *stse* processed by the SERNR postfilter is more similar to the original one than the remaining approaches. It is more

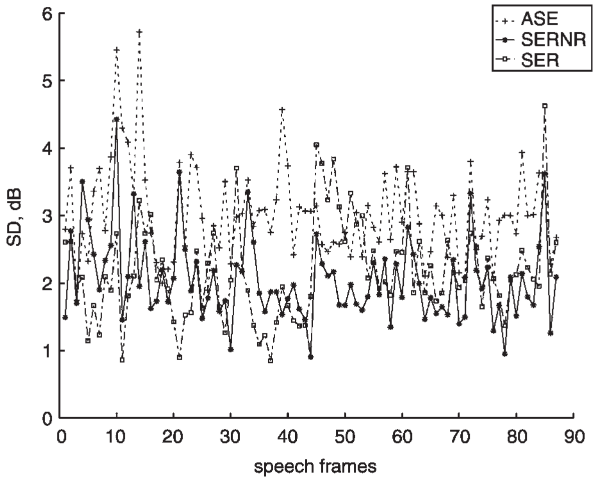


Fig. 8 *Spectral distance (SD) versus speech frames of utterance processed by three different postfiltering techniques*

effective in restoring the *stse* and reducing the coding noise of the processed speech. It should be remarked that although the SER postfilter presents low SD values at a few points, it does not attack the coding noise as the SERNR postfilter does. Indeed, from informal listening tests we perceived that the SERNR method is capable of considerably improving the quality of decoded speech and is superior to the ASE and SER techniques. Further results on the performance of the SERNR postfilter — including the ITU-T P. 862 perceptual evaluation of speech quality (PESQ) standard — are presented and discussed in Section 6.

4.2 Noise suppression

Noise suppression techniques have become of paramount importance in very low bit rate speech coding applications, because they can reduce background noise and enhance the quality of synthesised speech [24, 27]. In this work, we employ a smoothed spectral subtraction (SSS) method [27] that performs significantly better than conventional spectral subtraction approaches. To describe the SSS technique, we assume a clean speech signal $s(t)$ and a stationary and uncorrelated additive white Gaussian noise $n(t)$. The power spectrum $P_y(\varpi)$ of the noisy speech $y(t) = s(t) + n(t)$ corresponds to the sum of the power spectra $P_s(\varpi)$ and $P_n(\varpi)$ of $s(t)$ and $n(t)$, i.e.

$$P_y(\varpi) = P_s(\varpi) + P_n(\varpi) \tag{19}$$

In the traditional spectral subtraction method described in [27], the estimated spectrum of the clean speech signal $\tilde{P}_s(\varpi)$ can be obtained by subtracting the estimated noise spectrum $\tilde{P}_n(\varpi)$ from the noisy speech spectrum, as given by

$$\tilde{P}_s(\varpi) = P_y(\varpi) - \tilde{P}_n(\varpi) \tag{20}$$

The spectral subtraction principle can be interpreted as a time-varying linear filter by using the Fourier transform and rewriting (20). After noise suppression is applied, the clean speech signal is given by

$$\begin{aligned}\hat{s}(t) &= F^{-1}\left\{\sqrt{\frac{P_y(\varpi) - P_n(\varpi)}{P_y(\varpi)}}Y(\varpi)\right\} \\ &= F^{-1}\{H(\varpi)Y(\varpi)\} = F^{-1}\{\hat{S}(\varpi)\}\end{aligned} \tag{21}$$

where $Y(\varpi)$ is the Fourier transform of the noisy speech, $H(\varpi)$ is a time-varying linear filter and $\hat{S}(\varpi)$ is the estimate of the Fourier transform of the clean speech. According to (21), spectral subtraction corresponds to a frequency dependent attenuation to each frequency in the noisy speech spectrum $P_y(\omega)$, where the attenuation varies with $\frac{P_y(\omega)}{P_n(\omega)}$ [24].

The SSS technique [27] involves three additional strategies over the traditional spectral subtraction. First, the $H(\omega)$ filter attenuation is limited to -10 dB, avoiding signal distortion. Second, the noise estimation is artificially increased by 5 dB in order to prevent speech deterioration when the noise spectrum is not properly estimated. And finally, we use smoothed versions of the FFT derived noisy speech and noise spectra estimates via a smoothing window [27], preventing the arise of musical noise.

5 Coder structure

The encoder and decoder schematics of the 1.2 kb/s codec are shown in Figs. 9 and 10, respectively. Following the encoder block diagram, after LP analysis has been performed on a 20 ms speech frame, the pitch detection algorithm described in Section 2.1 is invoked in order to locate any evidence of voicing. The LPC coefficients are transformed into LSF parameters and encoded with 21 bits per frame by the optimised switched-predictive vector quantiser presented in Section 3, the gain is uniformly quantised with 5 bits per frame and the excitation is encoded with 3 bits per frame. Speech frames classified as voiced are encoded by an MMBE structure, which was described in Section 2.2, while unvoiced frames are encoded by the modelling and synthesis technique introduced in Section 2.3. Voiced frames are split into three frequency bands, which are implemented with fixed filter banks, and bandpass voicing analysis is performed, as detailed in Section 2.2. The bit allocation for the analysed codec is shown in Table 4. In order to determine the average bit rate of the codec, a statistical analysis was conducted, comprising 4 min of speech. This speech material consists of 10 phonetically balanced sentences [25] (in the Portuguese spoken in Rio de Janeiro) uttered by four different speakers (two male and two female). This analysis has shown that our algorithm operates at an average rate of 1.2 kb/s.

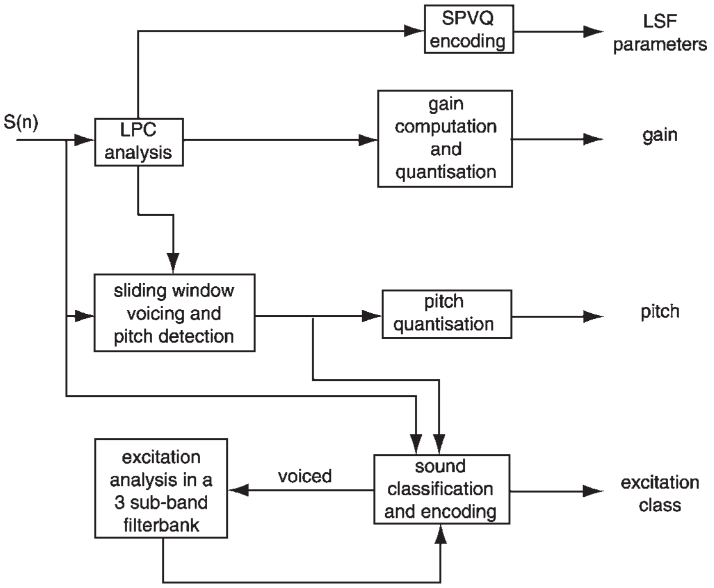

SPVQ encoding
LSF parameters
S(n)
LPC analysis
gain computation and quantisation
gain
sliding window voicing and pitch detection
pitch quantisation
pitch
excitation analysis in a 3 sub-band filterbank
voiced
sound classification and encoding
excitation class

Fig. 9 *Block diagram of 1.2 kb/s encoder*

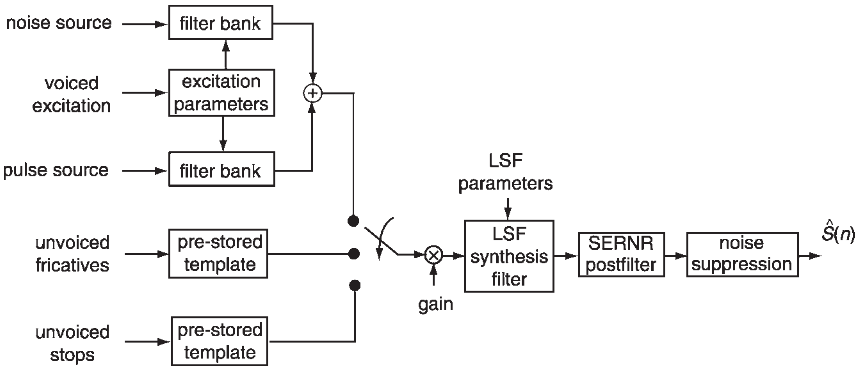


Fig. 10 *Block diagram of 1.2 kb/s decoder*

Table 4: Bit allocation

Parameters	Voiced	Unvoiced
LSFs	21	0
Gain	5	5
Excitation	3	3
Pitch	6	0
Total bits/20 ms	35	8
Bit rate	1.75 kb/s	0.4 kb/s

At the decoder, for voiced speech frames a pair of filter banks is used to generate the mixed excitation. The filter bank excitation is declared fully unvoiced for unvoiced frames and hence no voiced excitation is created. For the voiced frames, mixed excitation is generated as the sum of the filtered pulse and noise excitation, as described in Section 2.1. The next step is to perform the LPC synthesis with the coefficients corresponding to the interpolated LSFs and apply the decoded gain to the synthesised speech. The SERNR filter and the noise suppression technique, shown in Section 4, are then applied to the synthesised speech.

6 Objective and subjective tests results

We have carried out a number of preliminary informal assessments, in addition to A/B comparison and mean opinion scores (MOS) listening tests. We have also measured the spectral distortion (SD) of LSF quantisation and the perceptual evaluation of speech quality (PESQ) scores. The ITU-T PESQ standard is a valuable assessment tool because it provides strong correlation with subjective MOS scores and can easily be used by other researchers for verification purposes. All the results presented in this section are used in order to separately evaluate the improvements provided by each strategy presented in this paper, which are then incorporated in the 1.2 kb/s codec described in this paper.

The test material included both clean and noisy speech. According to preliminary informal listening tests conducted on clean speech, as well as on noisy speech at different SNR levels, the 1.2 kb/s codec was superior to the 2.4 kb/s MELP standard in tandem connections and comparable in non-tandem situations. These tests indicate that the preference for the 1.2 kb/s codec in clean speech was not altered for noisy speech. For these reasons, we show here only the results for clean speech. Ten (10) sentences were taken from lists of phonetically balanced sentences [25] (in the Portuguese spoken in Rio de Janeiro) and were uttered by 10 speakers (five male and five female). In the A/B comparison tests, the speech material was presented to 20 listeners and consisted of 10 sentence pairs, where each was uttered by a different speaker and processed by the 1.2 kb/s coder and the MELP. Since a particular sentence pair was also randomly presented in reverse order, there are 400 opinions for each test. In the preliminary tests reported in Sections 6.1 and 6.2 we have used a simpler MMBE speech coder platform, without noise suppression, using a standard pitch detection algorithm (without the improved method described in Section 2.1), and employing the SPVQ2 LSF quantiser without codebook optimisation.

6.1 Preliminary tests: stops and fricatives encoding

First we evaluated the stops and fricatives modelling and synthesis techniques and compared them with noise excitation. Two scenarios were considered for stops and fricatives modelling and synthesis techniques: an approach with pre-stored templates and a method using transmission of LPC sets. For comparison of the former with the noise excitation, the results show that 54% of listeners had no clear preference, 26% preferred the fricatives and stops approach, while 20% found the noise excitation based coder superior. This means that the stops and fricatives modelling and synthesis techniques perform slightly better than the noise excitation, used in the MELP standard, while it significantly reduces the bit rate. Comparison of the method which transmits LPC sets with noise excitation, show that 40% of listeners had no clear preference, 36% preferred the fricatives and stops approach, while 24% found the noise excitation based coder superior. This reveals that the fricatives and stops encoding scheme operating at a higher transmission rate performs better than noise excitation. In addition to the A/B comparison tests, we obtained PESQ scores for the same speech material. The scores were 2.37, 2.39 and 2.47 for noise excitation, fricatives and stops encoding without and with the transmission of LPC sets, respectively, confirming the preferences found in the A/B comparison tests. It is worth noting that fricatives and stops encoding without the transmission of LPC sets offers a very attractive trade-off between speech quality and bit rate and for this reason it has been chosen for use in the codec described in Section 5.

6.2 Preliminary tests: tandem connections and postfiltering evaluation

In tandem configurations the LSF quantisation is affected at each transcoding step. Certainly, this becomes more serious for quantisation procedures that provide worse performance. In this context, the 1.2 kb/s variable rate codec has the advantage of yielding better *stse* quantisation than the MELP algorithm (see Table 6 in Section 6). Furthermore,

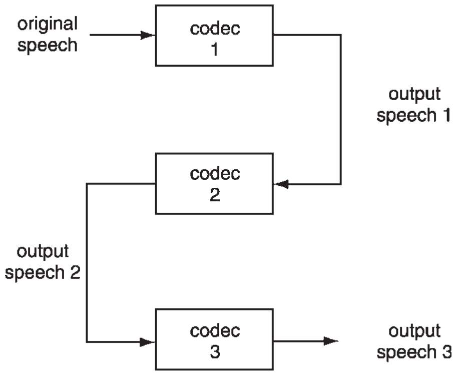


Fig. 11 *Speech codecs in tandem connections*

the ASE postfilter of Chen and Gersho [22] used in the MELP [3] attempts to suppress coding noise, but modifies the speech spectral envelope. In tandem connection situations this introduces distortion, which increases with the number of times the speech signal is encoded and decoded, and is reflected in the resulting speech quality. On the other hand, the SERNR [9, 23] approach employed in the 1.2 kb/s coder does not introduce this type of signal distortion, representing a more attractive choice in these situations. Our analysis is mainly focused on the effects of these connections on the perceptual quality of decoded speech, which can be estimated by the ITUT-P.862 PESQ recommendation [28]. However, subjective assessments through A/B comparison tests and LSF quantiser performance are also provided to corroborate the PESQ results. A block diagram of the speech codecs in tandem connections and the corresponding output decoded speech at each transcoding stage is depicted in Fig. 11. Note that in terms of LSF quantiser performance, the SD is computed between the original LSF parameters and the quantised ones, for no tandem connection. In the situations of one and two tandem connections, the SD is calculated between the original LSFs and the quantised LSFs by codecs 2 and 3, respectively.

In a preliminary experiment, we have compared the SERNR postfilter and the ASE filter followed by a fixed pulse dispersion (PD) filter, used in the MELP standard in one, two and no tandem connections. Note that we have left the SER postfilter out of this assessment because informal listening tests had indicated that this approach has a performance similar to the ASE postfilter. The SERNR postfilter was found to be superior by 37.5% of the listeners, while 18.2% showed a preference for the ASE filter followed by the PD filter, and 44.3% of them showed no preference. Then, we carried out a comparison of these postfilters in one tandem connection. The SERNR method was found to be superior by 47.8% of the listeners, whereas 12.5% showed a preference for the ASE and 39.8% of them had no preference. For two tandem connections, we found a 70.0% preference for the SERNR, 25.3% of listeners who had no preference and only 4.7% had preference for the ASE filter. Following the A/B comparison tests we obtained the PESQ scores for the same speech material. The scores are 2.39 and 2.32 for the SERNR and the ASE postfilters, respectively. For one tandem connection, we obtained 1.86 and 1.74 for the SERNR and the ASE techniques, respectively, whereas with two tandem connections the scores were 1.62 and 1.43 for the SERNR and the ASE techniques, respectively. Next, we show in Table 5 the results in terms of SD and percentage of outliers for the LSF quantisation of speech material after being processed by identical codecs with SERNR and ASE postfilters. All these experiments and results obtained with the two postfilters indicate that the SERNR postfilter is preferable to the ASE technique.

Table 5: Effect of postfilters on the LSF quantisation results

Postfilter	Tandem	SD, dB	%2–4, dB	%> 4, dB
ASE	No	1.03	2.15	0
SERNR	No	1.03	2.15	0
ASE	One	3.40	68.81	23.82
SERNR	One	2.95	62.80	17.09
ASE	Two	4.86	29.33	69.31
SERNR	Two	4.22	45.99	49.92

The performance of the speech coder of Section 5, operating at an average rate of 1.2 kb/s, was then evaluated and compared with the MELP coder, operating at 2.4 kb/s, in terms of both objective and subjective listening assessments: SD and percentage of outliers of LSF quantisation, A/B comparison and mean opinion score (MOS) tests and the ITU-T P.862 perceptual evaluation of speech quality (PESQ) standard. We also performed an evaluation of both speech coders in tandem connections and the results for these experiments are described in the following subsections.

6.3 LSF quantisation results: proposed × MELP

In order to assess the effects of tandem connections on the spectral envelope of speech, we chose the spectral distance (SD) given by (15) between the original and quantised LPC spectral envelopes as the performance index. For no tandem connection, the SD is computed between the original LSF parameters and the quantised ones. In a transcoding scenario, the SD is calculated between the original LSFs and the quantised LSFs by the codec in one and two tandem connections, respectively, as detailed in Section 6.2. The VQ performance in terms of average SD and percentage of outliers between 2 and 4 dB, and above 4 dB is given in Table 6 for all speakers used in the experiments.

The results shown in Table 6 indicate that the proposed codec introduces a milder type of short-term spectral distortion than the MELP. In tandem connections we note that the values of SD obtained for the 1.2 kb/s codec are considerably lower than the ones computed for the MELP. We also observe that in tandem connections the percentage of outliers above 4 dB — corresponding to severe distortion — is significantly higher for the MELP coder. As will be mentioned in the discussion of the subjective tests, the superior performance of the 1.2 kb/s scheme in terms of the average SD and the percentage of outliers above 4 dB significantly contributes to provide a better speech quality. These results indicate that the reason for the superiority of the 1.2 kb/s codec in tandem connections is the combined use of the SPVQ system and the SERNR postfilter.

Table 6: LSF quantisation results

LSF VQ	Tandem	SD, dB	%2–4, dB	%> 4, dB
MELP	No	1.35	2.27	0
Proposed	No	0.95	1.42	0
MELP	One	3.55	64.33	26.51
Proposed	One	2.75	52.74	12.68
MELP	Two	4.93	23.04	71.25
Proposed	Two	3.82	39.78	38.42

Table 7: A/B comparison tests

Tandem connections	No	One	Two
Proposed coder (%)	35	46.8	52.8
Comparable quality (%)	32	28.4	19.2
MELP (%)	33	24.8	28

6.4 A/B comparison tests: proposed × MELP

First, the speech coder platform of Section 5, operating at 1.2 kb/s, was compared with the MELP coder, operating at 2.4 kb/s. We used the same speech material, listeners, speakers and test procedure previously described. The results have shown that 32% of the listeners had no clear preference, 35% of them preferred the 1.2 kb/s coder, while 33% preferred the MELP coder. Note that our codec performed slightly better than the MELP coder, although it operates at half the bit rate of the North American standard.

In another situation, comparison of our coding scheme against the MELP was carried out with one tandem connection. The 1.2 kb/s coder was found to be superior by 46.8% of the listeners, while 24.8% showed a preference for the MELP and 28.4% of them showed no preference. The results of the A/B comparison tests are summarised in Table 7. It is clear from these results that for one tandem connection, the coder structure of Section 5 is superior to the MELP, corroborating the results (in terms of SD) of the LSF quantisation in Section 6.3.

6.5 Mean opinion score (MOS) tests: proposed × MELP

In MOS tests, listeners are asked to rate a system on an absolute scale, usually ranging from 1 to 5. The quality scale ranges from excellent, for grade 5, to very annoying, for grade 1. A training speech database, including good and bad speech, was presented to all listeners before the test, in order to prepare listeners to assess the quality of the sentences. The same material presented in the A/B comparison tests was used for the MOS tests and presented to another panel of 20 listeners, giving a total of 200 opinions for each situation. We conducted the MOS tests for the original speech material in order to provide a benchmark for the speech processed by the 1.2 kb/s coder, the 2.4 kb/s MELP coder and then the speech processed under one tandem connection by the 1.2 kb/s coder and the MELP.

Results have shown that the original speech used as a benchmark scored 4.17, while speech processed by the 1.2 kb/s and the MELP coders scored 3.00 and 2.97, respectively. The MOS results are given in Table 8 along with a 95% confidence interval ($\pm\delta$). For one tandem connection the speech processed by the proposed and MELP coders scored 2.41 and 2.19, respectively. These results confirm that the proposed coder is comparable with the MELP coder and is clearly superior in tandem connection situations, as suggested by the results for the SD of LSF quantisation and A/B comparison tests in previous subsections. For two tandem connections, we did not perform the MOS test because the quality was considered poor. The reader is referred to the results obtained by LSF quantisation, A/B comparison tests and PESQ scores, which indicate that the speech quality of the proposed codec was clearly less affected than the MELP.

Table 8: MOS tests

Situation	MOS ± δ
Original speech	4.17 ± 0.02
1.2 kb/s coder	3.00 ± 0.02
MELP	2.97 ± 0.03
1.2 kb/s coder - one tandem	2.41 ± 0.03
MELP - one tandem	2.19 ± 0.03

Table 9: PESQ tests

Situation	PESQ score
Original speech	4.50
1.2 kb/s coder	2.70
MELP	2.69
1.2 kb/s coder - one tandem	2.24
MELP - one tandem	2.13
1.2 kb/s coder - two tandem	2.03
MELP - two tandem	1.82

6.6 ITU-T P.862 PESQ standard results: proposed × MELP

The recently adopted ITU-T P.862 perceptual evaluation of speech quality (PESQ) recommendation is an objective measurement technique for estimating subjective quality obtained in listening-only tests [28]. PESQ compares an original speech signal with a degraded signal that is the result of passing the original signal through a communications system. The output of PESQ is a prediction of the perceived quality that would be given to the decoded speech by listeners in a subjective listening test such as the MOS test. The PESQ score is mapped to a MOS-like scale with range between 1.0 and 4.5. In this section we assess and compare our codec to the MELP coder through PESQ tests.

The same material used for LSF quantisation, A/B comparison and the MOS tests was used for the PESQ assessment, whose scores are shown in Table 9. The results show that the original speech used as a benchmark scored 4.5, while the processed speech for the 1.2 kb/s and MELP coders scored 2.70 and 2.69, respectively. For one tandem connection the processed speech for the 1.2 kb/s and the MELP coders scored 2.24 and 2.13, respectively, whereas for two tandem connections our codec and the MELP scored 2.03 and 1.82, respectively. These results corroborate that the 1.2 kb/s coder is comparable to the MELP coder and is clearly superior in tandem connection situations.

7 Conclusions

Several techniques to improve the performance of very low bit rate speech coders have been presented, analysed and shown to be useful tools for this important application. We have also incorporated these techniques in a variable rate speech coder structure that operates at an average rate of 1.2 kb/s. An efficient modelling and synthesis technique is described, examined and employed to encode unvoiced frames at only 0.4 kb/s, showing an attractive trade-off between speech quality and bit rate. We adopted a pitch detection algorithm that uses a sliding window to further reduce incorrect pitch values and voicing decisions. We employed a classification algorithm that distinguishes voiced, unvoiced fricative, unvoiced stop and silence frames. We proposed an efficient LSF switched-predictive vector quantisation (SPVQ) structure with jointly optimised codebooks, which can save more than 4 bits per vector compared to MVQ (memoryless vector quantiser) without reduction in performance, under noiseless conditions. A spectral envelope reconstruction combined with noise reduction (SERNR) postfilter, that combines the strengths of the ASE and the SER strategies, is presented, analysed and used as an important strategy for speech quality enhancement. An evaluation of the SERNR postfilter has shown that it has a superior performance compared to the ASE and the SER techniques. We remark that the SERNR postfilter performance is closely related to the LSF quantiser performance because our approach attempts to reconstruct the *stse* obtained from the received LSF parameters. Therefore, using a high quality SPVQ encoding scheme makes possible an accurate *stse* restoration and can deliver higher quality decoded speech. The 1.2 kb/s speech coder described in this paper, that incorporates all these techniques, is comparable (slightly outperforming) the MELP coder, operating at 2.4 kb/s. This conclusion was obtained from the computations of spectral distortion of LSF quantisation and the application of A/B comparison, MOS and PESQ tests. For one tandem connection, the coder described and studied in detail in this paper is clearly superior to the MELP coder. Finally, we remark that most of the strategies presented in this paper can also be adopted in the design of other classes of speech coders, such as CELP structures recommended by the ITU, in order to improve speech quality.

8 References

1 Griffin, D.W., and Lim, J.S.: 'Multiband excitation vocoder', *IEEE Trans. Acoust. Speech Signal Process.*, 1988, pp. 1223–1235

2 Teague, K.A., Leach, B., and Andrews, W.: 'Development of a high-quality MBE based vocoder for implementation at 2400 bps'. Proc. IEEE Wichita Conf. on Communications, Networking and Signal Processing, April 1994, pp. 129–133

3 Supplee, L.M., Cohn, R.P., Collura, J.S., and McCree, A.V.: 'MELP: the new Federal Standard at 2400 bps'. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 1997, pp. 1591–1594

4 Schroeder, M.R., and Atal, B.S.: 'Code-excited linear prediction (CELP): high quality speech at very low bit rates'. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 1985, pp. 937–940

5 Atal, B., and Hanauer, S.: 'Speech analysis and synthesis by linear prediction of the speech wave', *J. Acoust. Soc. Am.*, 1971, **50**, pp. 637–655

6 de Lamare, R.C., and Alcaim, A.: 'Very low bit rate speech coding in tandem connections', *Electron. Lett.*, 2003, **39**, (18), pp. 1356–1357

7 Unno, T., Barnwell T.P., III, and Truong, K.: 'An improved mixed excitation linear prediction (MELP) coder'. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Phoenix, USA, 1999, paper 1764

8 Ehnert, W.: 'Variable-rate speech coding: coding unvoiced frames with 400 bps'. Proc. EUSIPCO'98, Rhodes, Greece, 1998, pp. 1437–1440

9 de Lamare, R.C., da Silva, L.M., and Alcaim, A.: 'Sound specific modelling and synthesis with a new postfiltering in low bit rate speech coding'. Proc. IEEE Int. Symp. on Circuits and Systems, Scottsdale, Arizona, USA, 2002

10 Soong, F.K., and Juang, B.H.: 'Line spectrum pair (LSP) and speech data compression'. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 1984

11 Paliwal, K.K., and Atal, B.S.: 'Efficient vector quantization of LPC parameters at 24 bits/frame', *IEEE Trans. Speech Audio Process*, 1993, **1**, (1), pp. 3–14

12 LeBlanc, W.P., Battacharya, B., Mahmoud, S.A., and Cupperman, V.: 'Efficient search and design procedures for robust multi-stage VQ of LPC parameters for 4 kb/s speech coding', *IEEE Trans. Speech Audio Process.*, 1993, **1**, (4), pp. 373–385

13 Gersho, A., and Gray, R.M.: 'Vector quantization and signal compression' (Kluwer Academic Publishers, Boston, 1992)

14 da Silva, L.M., and Alcaim, A.: 'Differential coding of speech LSF parameters using hybrid vector quantization and bidirectional prediction', *IEEE Trans. Speech Audio Process.*, 2000, **8**, (2), pp. 208–210

15 Yong, M., Davidsson, G., and Gersho, A.: 'Encoding of LPC spectral parameters using switched-adaptive interframe vector prediction'. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, New York, USA, 1988, Vol. 1, pp. 402–405

16 Eriksson, T., Lindén, J., and Skoglund, J.: 'Exploiting interframe correlation in spectral quantization: a study of different memory VQ schemes'. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 1996

17 McCree, A., and De Martin, J.C.: 'A 1.7 KB/S Melp coder with improved analysis and quantization'. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 1998

18 de Lamare, R.C., and Alcaim, A.: 'Analysis of LSF switched-predictive vector quantisers'. Proc. IEEE Int. Symp. on Signal Processing and its Applications, Kuala Lumpur, Malaysia, 2001

19 de Lamare, R.C., and Alcaim, A.: 'Noisy channel performance of LSF switched-predictive vector quantisers'. Proc. IEEE Int. Conf. on Information, Communications and Signal Processing, Singapore, 2001

20 Barnes, C.F., and Frost, R.L.: 'Vector quantizers with direct sum codebooks', *IEEE Trans. Inf. Theory*, 1993, **39**, pp. 565–580

21 Chen, J., and Gersho, A.: 'Adaptive postfiltering for quality enhancement of coded speech', *IEEE Trans. Speech Audio Process.*, 1995, **3**, pp. 59–71

22 da Silva, L.M., and Alcaim, A.: 'Enhancement of CELP speech coding with postfiltering for spectral envelope restoration'. Proc. Int. Conf. on Telecommunications, Porto Carras, Greece, June 1998, pp. 269–272

23 de Lamare, R.C., and Alcaim, A.: 'Effects of adaptive postfilters on the LSF quantisation for low bit rate speech coders in tandem connections', Proc. IEEE Int. Symp. on Signal Processing and its Applications, Paris, 2003

24 Arslan, L., McCree, A., and Viswanathan, V.: 'New methods for adaptive noise supression'. Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing, 1995, pp. 373–385

25 Alcaim, A., Solewicz, J.A., and Moraes, J.A.: 'Frequency of occurrence of phones and lists of phonetically balanced sentences in the Portuguese spoken in Rio de Janeiro', *J. Brazilian Telecommun. Soc.*, 1992, **7**, pp. 23–41

26 Laroia, R., Phamdo, N., and Farvardin, N.: 'Robust and efficient quantization of speech LSP parameters using structured vector quantizers'. Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing, Maio, 1991, pp. 641–644

27 Boll, S.F.: 'Suppression of acoustic noise in speech using spectral subtraction', *IEEE Trans. Acoust., Speech, Signal Process.*, 1979, **27**, pp. 113–120

28 'Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs', Recommendation P.862, ITU-T, February 2001