Analysis of Postfilters for Low Bit Rate Speech Coders in Tandem Connections

Rodrigo C. de Lamare and Abraham Alcaim
CETUC - PUC-RIO, 22453-900, Rio de Janeiro - Brazil
e-mails: delamare@infolink.com.br, alcaim@cetuc.puc-rio.br

Abstract—**In this paper we analyse postfiltering techniques for very low bit rate speech coders in tandem connections. A mixed multiband excitation (MMBE) linear predictive coding (LPC) algorithm, that encodes voiced frames at 1.75 kb/s and unvoiced frames at 0.4 kb/s, is employed to assess the performance of different postfilters in tandem connections. We perform a comparative analysis of the well known adaptive spectral enhancement (ASE) technique with a recently reported approach, called spectral envelope restoration combined with noise reduction (SERNR) postfilter, using the same MMBE platform. Subjective listening tests in tandem connections show that the SERNR technique is clearly superior to the ASE postfilter.**

I. Introduction

Postfiltering techniques are paramount in low bit rate speech coding algorithms because they can enhance the quality of synthesised speech, mitigating some of the effects that degrade its subjective quality such as artifacts and harshness. With new applications such as voice over IP networks (VOIP), very low bit rate speech coding algorithms have taken an increased importance and the role of adaptive postfilters has become fundamental, since they can contribute with better subjective quality to the decoded speech and do not require additional transmitted bits. Most modern very low bit rate speech coding algorithms such as the mixed multiband excitation (MMBE) [1] and the mixed excitation linear prediction (MELP) [2] 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 [3]. The speech codec platform treated in this paper is based on an improved MMBE system, that employs a switched predictive vector quantiser technique (SPVQ) [4] to encode the LSF parameters and a sound specific modelling and synthesis approach to encode non-stationary sounds. To encode voiced frames, an MMBE approach with three sub-bands is used, whilst fricatives and stops modelling and synthesis techniques are used for unvoiced frames [5]-[7]. To reduce coding noise and improve decoded speech, a postfiltering technique is usually used at the front end of the codec. One of the most popular and successful postfiltering techniques, the adaptive spectral enhancement (ASE) [8] postfilter, is compared with a recently reported approach, the spectral envelope restoration combined with noise reduction (SERNR) [7] postfilter, using the same MMBE platform.

In digital telephony, it is often necessary to encode and decode speech signals more than one time, resulting in speech deterioration. In this work, the level of speech degradation imposed to an MMBE speech compression algorithm employing the ASE and the SERNR postfiltering techniques is assessed in such practical situations.

This paper is organised as follows. Section II gives a general overview of the MMBE coder platform and its components. Section III discusses the adaptive postfilters structures. Section IV describes the tandem connection situations investigated in this work and Section V presents the results of subjective listening tests. Finally, Section VI summarises the main conclusions of this work.

II. Overview of the MMBE Coder Platform

Low bit rate speech coders that follow the classical vocoder principle of Atal and Hanauer [3] usually result in synthetic speech quality due to an impairment generally termed as ´buzziness´. Mixed Multiband Excitation (MMBE) [1],[2] addresses the problem of ´buzziness´directly, 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 sub-band in the speech frame.

A. Overview

The encoder and decoder schematics of the MMBE codec are shown in Fig. 1, respectively. Following the encoder schematic, after LP analysis has been performed on a 20 ms speech frame, a pitch detection algorithm similar to the one employed in the MELP [2] 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 a switched-predictive vector quantiser [4], the gain is quantised with 5 bits per frame and the excitation is encoded with 3 bits per frame. Speech frames classified as voiced are split into 3 frequency bands, which are implemented with fixed filter banks, and a bandpass voicing analysis is performed. For unvoiced frames, we use a modelling and synthesis approach that is described in [5],[6]. The bit allocation of the speech coder used in this work is shown in Table 1. The average bit rate of this codec is 1.2 kb/s.

At the decoder, the voiced speech frames are filtered by a pair of filter banks. For the voiced frames, mixed excitation is generated as the sum of the filtered pulse and noise excitation. 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. An adaptive spectral enhancement (ASE) filter followed by a pulse dispersion (PD) filter or a SERNR filter are then

Table 1: 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

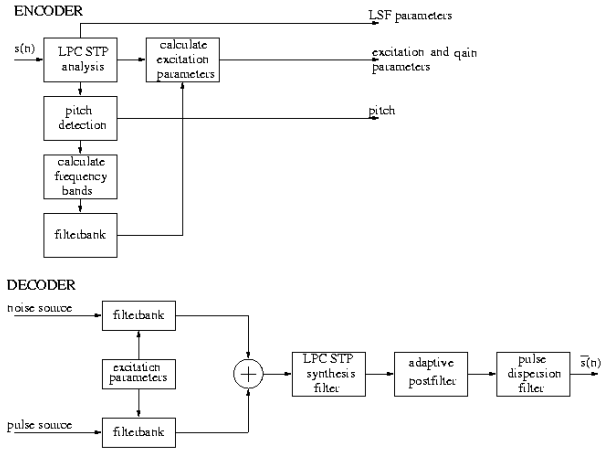

Fig. 1. Block diagram of the MMBE encoder and decoder.

proposed coder. This LSF quantiser combines memoryless vector quantisation (MVQ) and predictive vector quantisation (PVQ) 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, is encoded and transmitted [4]. The SPVQ system employed in this work is shown in Fig. 2. It operates at 21 bits per frame and uses 2 tree-structured multistage vector quantisers (1 PVQ and 1 MVQ). The SPVQ performance in terms of average spectral distortion and percentage of outliers between 2 and 4 dB, and above 4 dB is given in Table 2.

Table 2: Performance of the SPVQ system.

$SD(dB)$	1.0219
% 2 - 4 dB	2.5931
% $>$ 4 dB	0

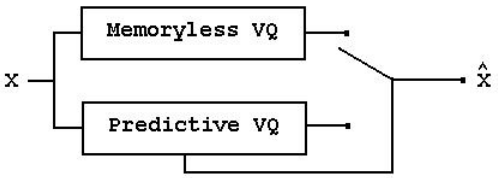

Fig. 2. SPVQ using 1 PVQ and 1 MVQ schemes.

applied to the synthesised signal.

B. Fricatives and Stops Encoding

In order to provide a clearer speech quality for the sentences containing stop and fricative sounds, we use an strategy based upon the algorithms introduced by Unno *et al.* [5] and Ehnert [6]. It envolves the detection and the modelling and synthesis of these signals.

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 maximizes the peakiness value [5]. In our approach there are two types of stop signals since two excitation codebook entries are reserved for these sounds. The detection of fricative sounds makes use of appropriate thresholds for the zero crossings and the energy of each frame.

All stop and fricative signals $f|s(n)$ are produced by scaling and LPC filtering pre-stored templates of LPC residual signals $r(n)$ using templates of LPC coefficients. The templates are carefully chosen to avoid the transmission of the LPC coefficients for unvoiced frames. We have used one residual signal and an LPC set as templates to synthesise fricatives, whilst two residual signals and two LPC sets were employed to reproduce stops [7].

C. LSF Quantisation

An efficient LSF quantisation scheme denoted switched predictive vector quantisation (SPVQ) [4] is adopted in the

III. Adaptive Postfilters

One strategy to reduce the perceived coding noise makes use of an adaptive postfilter at the output of the speech decoder. The adaptive spectral enhancement (ASE) postfilter [8] is the most usual and popular structure and has the following transfer function:

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

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 ν 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 [8]. In the MELP standard this filter is followed by a fixed pulse dispersion filter (PD) [2], based on a spectrally-flattened triangle pulse, that spreads the excitation energy within a pitch period, reducing 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) [9] and has the following transfer function:

$$H_{SER} = \frac{\tilde{A}(z/\xi)}{A(z/\xi)} \tag{2}$$

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.

A recently reported strategy [7] to enhance the quality of decoded speech combines the strengths of the ASE and the spectral envelope restoration (SER) [9] postfilters. This structure, called spectral envelope reconstruction and noise reduction (SERNR) postfilter [7], gathers the *stse* restoration properties of the SER filter and noise reduction capabilities of the ASE technique. The SERNR postfilter [7] has the following transfer function:

$$H_{SERNR} = \frac{\tilde{A}(z/\zeta)}{A(z/\eta)}(1 - \nu z^{-1}) \qquad (3)$$

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 ν are 0.82, 0.9 and $0.3k_1$, respectively. Note that the SERNR postfilter performance is closely related to the LSF quantiser performance because it attempts 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 coding structure described in [4] is used.

In Fig. 3 the *stse* of a speech segment is shown for the original speech, the encoded speech with the ASE filter followed by the PD filter, the SER postfilter and the SERNR postfilter. Note that the *stse* processed by the SERNR postfilter is more similar to the original one than the remaining approaches. It is superior in restoring the *stse* and reducing the coding noise of the processed speech. Indeed, from informal listening tests we have perceived that the SERNR method is capable of considerably improving the quality of decoded speech and is superior to the ASE and SER techniques.

IV. Tandem Connections

In digital telephony applications, it is often necessary to encode and decode speech signals more than one time, as depicted in Fig. 4. These situations are called tandem connections and usually result in some speech degradation, because of the inherent losses caused by the compression techniques. In this work, we assess the performance of the ASE and the SERNR postfiltering techniques operating with the MMBE speech codec described in Section 2, in one and two tandem connections.

In the MELP coder [2], for instance, the ASE postfilter 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. On the other hand, the SERNR [7] approach does not introduce this type of signal distortion, representing a more attractive

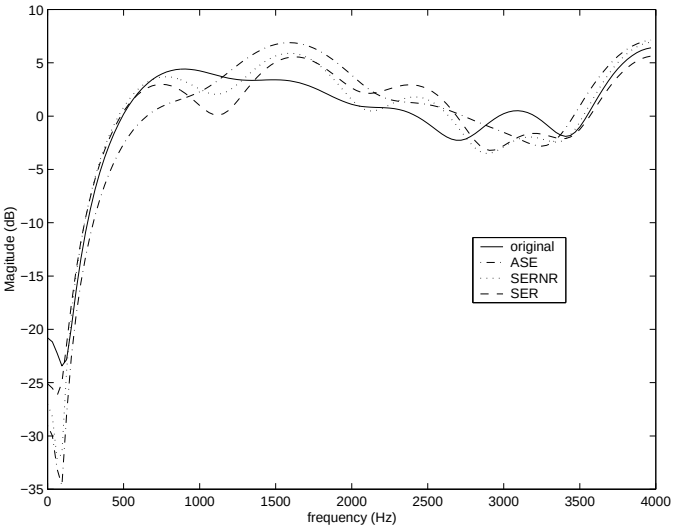

Fig. 3. The *stse* of a speech segment processed by three different postfiltering techniques.

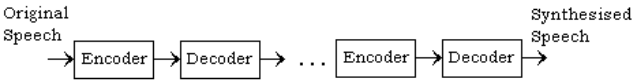

Fig. 4. Block diagram of the tandem connections.

choice in these situations, as will be shown in the next section.

V. Performance Results

To evaluate the performance and compare the ASE and the SERNR postfiltering techniques in one, two and no tandem connections using the MMBE speech coding platform described in section II, we conducted three independent A/B comparison tests with 10 sentence pairs, where each was uttered by a different speaker. Five female and five male speakers were used in the experiments. The test material included only clean speech and was presented to 20 listeners. Since a particular sentence pair was also randomly presented in reverse order, there are 400 opinions for each test.

In the first situation, the postfilters were compared with no tandem connections. The results, depicted in Fig. 5, have shown that 37% of the listeners preferred the SERNR postfilter, 19% found that the ASE technique was superior, whilst 44% had no clear preference.

In the second situation, a comparison of the SERNR postfilter against the ASE was carried out in one tandem connection. The SERNR method was found to be superior by 48% of the listeners, whereas 12% showed a preference for the ASE and 40% of them had no preference, as shown in Fig. 6.

In the third situation, two tandem connections were considered. The SERNR was preferred by 70% of the listeners, whilst only 5% found the ASE superior and 25% of them

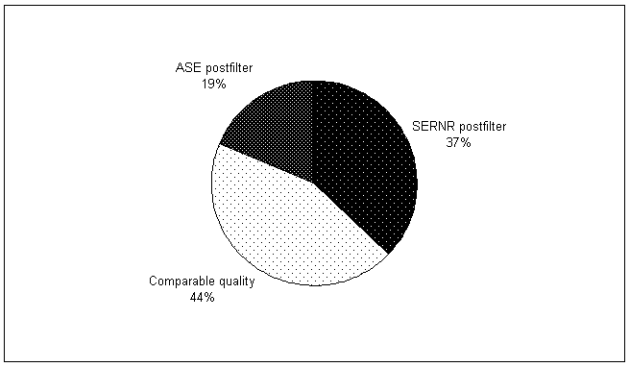

Fig. 5. A/B Comparison tests results.

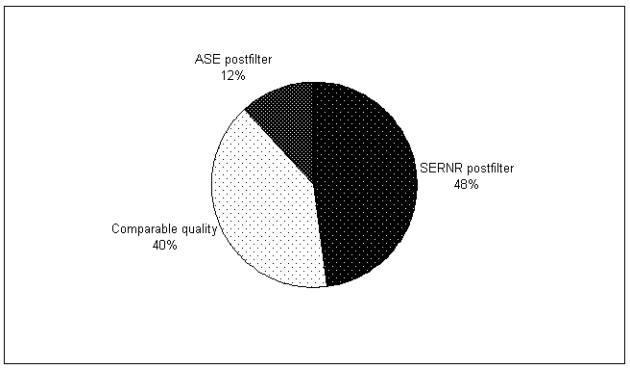

Fig. 6. A/B Comparison tests results in one tandem connection.

had no preference, as shown in Fig. 7. The results of the A/B comparison tests are shown in Table 3. It is clear from these results that in tandem connections the SERNR technique is definitely superior to the ASE approach.

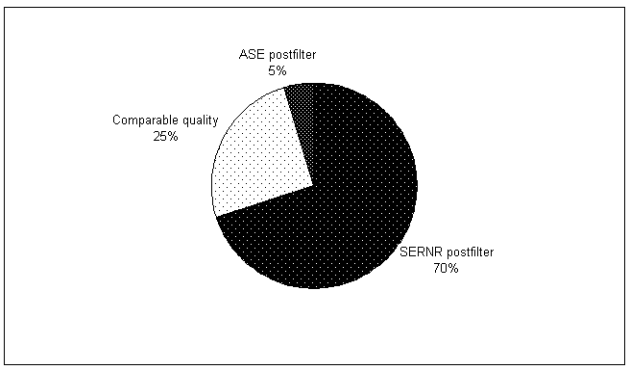

Fig. 7. A/B Comparison tests results in two tandem connection.

VI. Conclusions

We have conducted a performance evaluation of post-filtering techniques for a Mixed Multiband Excitation (MMBE) very low bit rate speech coder in tandem connections. The spectral envelope reconstruction and noise reduction (SERNR) postfiltering technique was compared to the traditional adaptive spectral enhancement (ASE) postfilter using an MMBE speech coding platform in one, two and no tandem connections. Subjective listening tests have shown that the SERNR postfilter is definitely superior to the ASE approach, in one, two and no tandem connections, at the expense of a higher computational complexity.

Table 3
A/B comparison tests.

Tandem connections	No	One	Two
SERNR postfilter (%)	37	48	70
Comparable quality (%)	44	40	25
ASE postfilter (%)	19	12	5

References

[1] K. A. Teague, B. Leach and W. Andrews, "Development of a high-quality MBE based vocoder for implementation at 2400 bps", *Proc. IEEE Wichita Conf. Communications, Networking and Signal Processing*, pp. 129-133, 1994.

[2] L. M. Supplee, R. P. Cohn, J. S. Collura and A. V. McCree, "MELP: The New Federal Standard at 2400 bps", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 1591-1594, 1997.

[3] B. Atal, S. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave", *The Journal of the Acoustical Society of America*, 1971, vol. 50, no. 2, pp. 637-655.

[4] R. C. de Lamare and A. Alcaim, "Analysis of LSF Switched-Predictive Vector Quantisers", *Proc. International Symposium on Signal Processing and its Applications*, Kuala-Lumpur, Malaysia, 2001.

[5] T. Unno, T. P.Barnwell III and K. Truong, "An Improved Mixed Excitation Linear Prediction (MELP) Coder", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Phoenix, USA, 1999

[6] W. Ehnert, "Variable-rate speech coding: coding unvoiced frames with 400bps", *Proc. EUSIPCO'98*, Rhodes, Greece, pp. 1437-1440, 1998.

[7] R. C. de Lamare, L. M. da Silva and A. Alcaim, "Sound specific modelling and synthesis with a new postfiltering in low bit rate speech coding", *Proc. IEEE International Symposium on Circuits and Systems*, Scottsdale, Arizona, USA, 2002.

[8] J. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech", *IEEE Trans. Speech and Audio Processing*, 1995, vol. 3, pp. 59-71.

[9] L. M. da Silva and A. Alcaim, "Enhancement of CELP Speech Coding with Postfiltering for Spectral Envelope Restoration", *Proc. ICT'98*, Porto Carras, Greece, pp. 269-272, June 1998.