Comparison of a Low Bit Rate Speech Compression Algorithm with the MELP Standard for Fricatives and Stops

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 examine the fricatives and stops encoding of a very low bit rate speech compression algorithm, based on a mixed multiband excitation system. The algorithm incorporates several improvements over previously reported coders. One of them is the use of a specific modelling and synthesis strategy for fricatives and stops at 400 b/s. The codec, which operates at an average rate of 1.2 kb/s, is compared with the North American standard 2.4 kb/s MELP coder for sentences with a large concentration of fricatives and stops sounds. Subjective listening tests indicate that the two codecs are comparable for both types of sounds, although the sound specific scheme operates at 400 b/s, whilst the MELP operates at 2.4 kb/s.

1 Introduction

The mixed multiband excitation (MMBE) [1] and the mixed excitation linear predictive (MELP) [2] coders are amongst the most modern and successful very low bit rate speech coding algorithms. These platforms 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. The use of the classical vocoder principle of Atal and Hanauer [3] often results in synthetic speech quality. The MELP and MMBE coders are able to improve the encoded speech quality by splitting the speech into several frequency bands [1],[2]. However, the mixed noise and pulse excitation is not capable of reproducing some specific signals such as those seen in stops and fricatives. In order to provide a clearer speech quality for the sentences containing these sounds, we use a strategy based on the algorithms introduced by Unno *et al.* [4] and Ehnert [5]. To transmit the spectral envelope information, we have chosen an LSF switched-predictive vector quantiser [6] that requires only 21 bits per frame to encode the LSF parameters. To reduce coding noise, a spectral envelope reconstruction and noise reduction (SERNR) postfilter is used. We conduct sound specific listening tests to compare the proposed algorithm with the MELP coder, for fricatives and stops sounds.

This paper is organised as follows. Section 2 briefly describes the speech compression algorithm. Fricatives and stops modelling and synthesis techniques are detailed in Section 3. Section 4 presents and discusses the results of subjective listening tests. Finally, Section 5 gives the concluding remarks.

2 Overview of the Compression Algorithm

The speech compression algorithm is based on a Mixed Multiband Excitation LPC [1] system. The speech input is split into several frequency bands, with a voiced excitation source or an unvoiced excitation source for each sub-band in the speech frame.

After LP analysis has been performed on a 20 ms speech frame, a robust pitch detection algorithm similar to the one employed in [5] is invoked to locate any evidence of voicing. The LPC coefficients are transformed into LSF parameters and encoded with 21 bits per frame by an optimised switched-predictive vector quantiser [6]. 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 specific modelling and synthesis approaches based on the techniques described in [4] and [5]. Voiced speech is encoded at 1.75 kb/s using an MMBE model, whilst unvoiced speech is encoded at 400 b/s, with the scheme described in Section 3.

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 SERNR postfilter [7] and a noise suppression method [8] are then applied to the synthesised signal.

The strategy to enhance the quality of decoded speech combines the strenghts of the adaptive spectral enhancement (ASE) [9] and the spectral envelope restoration (SER) [10] postfilters. This structure, called spectral envelope reconstruction and noise reduction postfilter (SERNR) [7], gathers the *stse* restoration properties of the SER filter and noise reduction capabilities of the ASE technique. The SERNR postfilter has the following transfer function:

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

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.

3 Fricatives and Stops Encoding

The mixed excitation allows the MMBE to have considerable freedom for the voicing decision. However, the mixed noise and pulse excitation is not capable of reproducing specific signals such as those seen in stops and fricatives. In order to provide a clearer speech quality for the sentences containing these sounds, we use an strategy based on the algorithms introduced by Unno *et al.* [4] and Ehnert [5]. It envolves the detection and the modelling and synthesis of these signals.

For the detection of stops 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 [4]. In our 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 whilst the second one is associated to those whose maximum amplitudes are found in the second part.

The detection of fricatives 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 whilst voiced frames typically do not cross the axis more than 60 times per frame [5].

In this model, all stop and fricative signals $f|s(n)$ are produced by scaling and LPC filtering prestored templates of LPC residual signals $r(n)$ using templates of LPC coefficients:

$$f|s(n) = Gr(n) + \sum_{i=1}^{p} a_i f|s(n-i) \tag{2}$$

where G is the gain based on the energy of the input stop or fricative signal and $a_1, ..., a_p$ are the LPC coefficients stored in the decoder. 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. The fricative and stop sounds are reproduced by the application of (2), where the residual signals and LPC coefficients templates are used with the transmitted gains for the synthesis of these sounds.

4 Subjective Tests Results

To evaluate the quality of the synthesised fricatives and stops, two independent A/B comparison tests were carried out. We have chosen two sentences in Brazilian Portuguese with a large concentration of fricatives and stops. The first sentence **("Vi Zé fazer essas viagens seis vezes")** contains a high proportion of fricative sounds, whereas the second one **("O atabaque do Tito é coberto com pele de gato")** has a large concentration of stops. Each sentence was uttered by five female and five male speakers. The test material included only clean speech and was presented to 10 listeners. Since a particular sentence was also randomly presented in reverse order, there are 200 opinions for each test.

In the first situation, we compared our platform with the MELP coder for speech with a large concentration of fricative sounds. The results have shown that 30% of the listeners preferred the proposed system, 29% found the MELP coder superior, whilst 41% had no clear preference. These opinions indicate that for fricative sounds, the proposed compression algorithm is comparable to the MELP.

In the second situation, a similar comparison was carried out for the speech material with a large concentration of stop signals. The MELP was found to be superior by 31% of the listeners, 29% preferred the proposed algorithm, whilst 40% had no clear preference, as shown in Table 3. For stop signals, the listening tests have shown that the proposed algorithm is also comparable to the MELP. Indeed, the results of the A/B comparison tests indicate that the proposed system, operating at 1.2 kb/s and encoding fricative and stop sounds with only 400 b/s has, for these types of sounds, a subjective quality comparable to the MELP standard operating at 2.4 kb/s.

Table 3
A/B comparison tests for stop and fricative sounds.

Sentences	Fricatives	Stops
Proposed(%)	30	29
Comparable(%)	41	40
MELP(%)	29	31

5 Concluding Remarks

We have performed a comparative analysis of a very low bit rate speech compression platform using sound specific techniques with the North American standard MELP coder. The encoding algorithm is based on several improvements (such as an efficient LSF quantisation, fricatives and stops modelling and synthesis techniques and the SERNR postfiltering) over previously reported mixed multiband excitation (MMBE) speech coders. We conducted a comparison of the proposed coder, operating at 1.2 kb/s and encoding fricative and stop sounds with only 400 b/s, with the MELP coder, operating at 2.4 kb/s, for fricative and stop sounds. Subjective listening tests indicate that, for these sounds, the proposed coder is comparable to the MELP, whilst it encodes unvoiced and silence frames with only 400 b/s.

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] 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

[5] W. Ehnert, “Variable-rate speech coding: coding unvoiced frames with 400bps”, *Proc. EUSIPCO’98*, Rhodes, Greece, pp. 1437-1440, 1998.

[6] 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.

[7] R. C. de Lamare, L. M. da Silva and A. Alcaim, “Fricatives and Stops Modelling and Synthesis with Improved Postfiltering in MMBE Coders”, *Proc. IEEE International Symposium on Circuits and Systems*, Scottsdale, Arizona, USA, 2002, accepted for publication.

[8] L. Arslan, A. McCree and V. Viswanathan, “New Methods for Adaptive Noise Supression”, *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 373-385, 1995.

[9] 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.

[10] 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, 1998.