Perceptual Evaluation of a new subband low bit rate speech compression system based on waveform Vector Quantization and SVD postfiltering

S.-E. Fotinea, I. Dologlou, S. Bakamidis, G. Stainhauer and G. Carayannis

Institute for Language and Speech Processing
Epidavrou & Artemidos 6, 151 25 Maroussi, Greece
evita@ilsp.gr

Abstract

This paper proposes a new low rate speech coding algorithm, based on a subband approach. At first, a frame of the incoming signal is fed to a low pass filter, thus yielding the low frequency (LF) part. By subtracting the latter from the incoming signal the high frequency (HF), non-smoothed part is obtained. The HF part is modeled using waveform vector quantisation (VQ), while the LF part is modeled using a spectral estimation method based on a Hankel matrix, its shift invariant property and SVD, called CSE. At the receiver side an adaptive postfiltering based on SVD is performed for the HF part, a simple resynthesis for the LF part, before the two components are added in order to produce the reconstructed signal. Progressive speech compression (variable degree of analysis/synthesis at transmitter/receiver) is thus possible resulting in a variable bit rate scheme. The new method is compared to the CELP algorithm at 4800 bps and is proven of similar quality, in terms of intelligibility and segmental SNR. Moreover, perceptual evaluation tests of the new method were conducted for different bit rates up to 1200 bps and the majority of the evaluators indicated that the technique provides intelligible reconstruction.

Introduction

A subband approach is first used to split the incoming signal into a Low Frequency (LF) and a High Frequency (HF) part. Efficient coding of the HF part may be achieved using quantization methods and in particular waveform vector quantization (VQ), while the LF part may be efficiently represented by means of models resulting in combinations of exponentially damped sinusoids. In this way, the whole approach provides a hierarchical compression scheme. A schematic view of the proposed system is presented in Figure 1. The next section provides detailed description of the different parts of the system. Our approach allows for progressive speech compression since it involves a variable degree of analysis/synthesis at the transeiver depending on the availability of channel capacity and/or the requirements of the specific application.

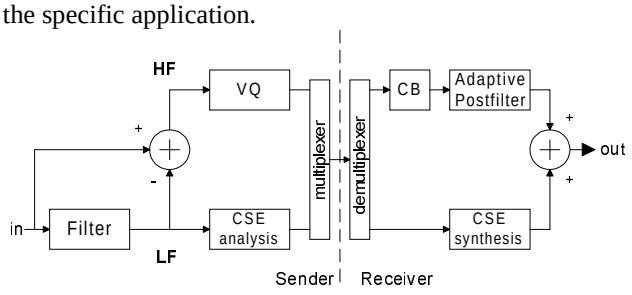

Figure 1: Schematic View of the new system

System Description

The algorithm processes the incoming signal on a frame by frame basis. First, a frame of the incoming signal is fed to a low pass filter. We use an 8-th order chebyshev low pass filter with 0.2 dB of ripple in the pass band and cut-off frequency *fco*. The resulting "linear"', LF signal serves as input to the CSE algorithm (Fotinea *et al*, 2000). The "non-linear" HF signal is obtained by subtracting the LF part from the incoming signal. This HF part is fed to the waveform VQ algorithm

Encoding

To encode the HF part of the incoming signal, we apply a straightforward waveform VQ algorithm, based on a code book (CB) generated by the well-known LBG algorithm (Gersho & Gray, 1992; Lemmerling, Dologlou & Van Huffel, 1998).

To encode the LF part, we use the CSE algorithm, with variable frame size and constant number *K* of complex damped exponentials to be modeled (for the work reported in this paper, K=6).

Each frame *lf_fr* of the LF signal s_{LF} is quantitated as a sum of six complex damped exponentials of the form:

$$s_{LF}(n) = \sum_{k=1}^{K} a_k e^{jp_k} e^{(-d_k + j2\pi f_k)n} =$$

$$\sum_{k=1}^{K} g_k z_k^n, n = 0,\ldots,N_s - 1 \quad (1)$$

Then, for each frame the CSE algorithm provides the estimates of the frequency f^k_fr, damping factor d^k_fr, amplitude a^k_fr and phases p^k_fr, *k=1,…,K*, which are then quantised.

It can be briefly mentioned at this point, that CSE uses the Hankel observation matrix *H* in order to construct its respective lower shift (top row deleted) and upper shift (bottom row deleted) equivalents (shift invariant properties), which are then truncated to order *K* by employing the SVD, resulting in matrices *Hv* and *H^* respectively.

Finally, calculation of a matrix *X =Hv pinv(H^)*, leads to frequency and damping factor estimates from the angles and eigenvalues respectively of the eigenvalues of *X*. Moreover, computation in a total least squares sense of estimates *g* follows, where the poles *z* are replaced by their estimates. In this way, the complex-valued liner parameter estimates, namely amplitude and phase are determined as the magnitudes and angles of *g* respectively.

Decoding

At the receiver side decoding of the HF part is performed by selection of the appropriate waveform from the CB. Afterwards an adaptive postfilter (Doclo *et al,* 1998) is applied to improve the perceptual quality of the reconstructed parts.

The postfiltering algorithm is based on a truncated singular value decomposition procedure, which had served in the past as tool for speech enhancement (Dendrinos, Bakamidis and Carayannis, 1991). The algorithm consists of two loops: the inner loop is an iterative procedure computing an enhanced signal from the receiver's input while the outer loop repeats the inner one a number of times, depending on the enhancement level. This algorithm can be seen in an alternative way as an adaptive FIR filtering operation. This postfiltering procedure cleans out the coarseness of the reconstructed HF signal, while it can be efficiently implemented, since it requires only the computation of the largest singular value and its corresponding singular vector.

Decoding of the LF part is achieved via a CSE based simple synthesis algorithm, where each frame of the LF part is resynthesized with the help of the Eq(1) by using its respective quantised frequency, damping factor, amplitude and phase estimates f^k_fr, d^k_fr, a^k_fr and p^k_fr, *k=1,…,K.*

Experimentation - Evaluation

In this section we compare our scheme to the CELP standard algorithm, applied to a speech signal sampled at 8 kHz, using 8 bits per sample. Our test signal consists of a phonetically balanced reference sentence, the latter containing 15000 samples (approximately 2 seconds of speech). The sentence is an enumeration of geographical places:

Paris, Bordeaux, Le Mans, Saint-Leu, Léon, Loudun.

For the CELP algorithm, we used a Fortran implementation of the Federal Standard 1016 4800 bps CELP vocoder (Campbell, Tremain and Welch, 1991).

In Figure 2, (a) the time domain representation of the first 2000 samples of the reference sentence, as well as its respective (b) High Frequency and (c) Low Frequency parts are depicted.

A brief discussion about the choice of different parameters in our scheme is presented below. The cut-off frequency *fco* plays the most important role in this scheme, since it determines the frequency content the VQ branch has to deal with. Moreover, parameters like the frame length used in CSE, can as well affect the bit rate that the proposed scheme achieves.

If *lfr* denotes the frame length used in CSE, *nsin* the number of sinusoids (*K*=2 *nsin*) used for coding of the LF signal part and we use, in average, 8 bits per sinusoidal parameter (namely frequency, damping factor, amplitude and phase), the bit rate of the LF coding part is given by:

$$d_{CSE} = \frac{4 * 8 * lfr \text{ bits}}{nsin \text{ samples}}$$

If $n = 2^w$ denotes the size of the code book CB of the HF part and N its respective dimension, the bit rate for the coding of the HF part is given by:

$$d_{VQ} = \frac{w \text{ bits}}{N \text{ samples}}$$

Finally, the bit rate for the overall system is calculated as follows:

$$d_{SYSTEM} = d_{CSE} + d_{VQ}$$

At first we designed the system at 4800 bps, in order to compare it with standard CELP.

With the following choice of parameters

$$f_{co} = 400Hz, nsin = 3, lfr = 500$$

$$n = 512 \ (w = 9), N = 21$$

we obtain a bit rate of 0.62 bits/sample or at a sampling frequency of 8kHz: 4964 Bit/sec. We thus, have a compression rate similar to CELP.

To assess the quality of the compressed speech, we use the following SNR definition:

$$SNR_{seg} \equiv 10\log_{10} \frac{1}{F} \sum_{j=1}^{F} \frac{\sum_{i=1}^{p} (in_j(i))^2}{\sum_{i=1}^{p} (in_j(i) - out_j(i))^2}$$

where *F* represents the number of frames, *p* is the frame length used in the computation, *in* the incoming signal and *out* the reconstructed one. In this respect, we calculate on per frame bases as follows:

$$in_j = in(1 + (j-1)p : jp)$$

$$out_j = out(1 + (j-1)p : jp)$$

For this experimentation *p* was chosen 500. The segmental SNRs for CELP and the new system were found to be: 11.69 DB and 12.26 dB respectively. It is worth mentioning that for CELP result, we made a comparison between the highpass filtered input and the non-postfiltered output (standard CELP applies at the end an adaptive postfilter routine to reduce perceptual coder noise). In this respect, we draw the conclusion that at comparable bit rates the new scheme obtains a quality similar to CELP. To further investigate this mathematical result, the following perceptual evaluation test was conducted.

Ten Greeks (six male and four female), with no hearing disability, with documented knowledge of French and unfamiliar whatsoever with this experimentation, were asked to evaluate the two sentences at the receiver side, namely the CELP and the new method's results. They were asked first to indicate whether both sentences were intelligible and then to mark which sentence was reconstructed in the best way. In Table 1, their answers are presented. All of them indicated the intelligibility of the two methods (100% intelligible). Despite the superiority of the new method over CELP in terms of SNR performance, CELP's quality is better preferred than that of the new method.

The new system can result in different bit rates by changing the parameters involved, i.e the number of sinusoids and frame length used for the LF part, as well as the code book length and dimension for the HF branch.

A new sentence was used for a perceptual evaluation of the new system for variable bit rate. Our second test signal contains 44000 samples (5.5 seconds of speech), and it is in Greek (in parenthesis the phonetic transcription as well as a translation in English can be found):

Είναι ένας ερευνητικός οργανισμός που δουλεύει σε θέματα σύνθεσης φωνής.
('ine 'enas erevnitik'os organism'os pu δul'evi se θ'emata s'inθesis fon'is / It is a research organization working on speech synthesis issues).
In order to achieve variable bit rates, we varied lfr for CSE and have created different code books. Below, the choices of all the parameters used in this experiment as well as the resulting bit rate are presented in tabular form (see Table 2).
In Figure 3, the magnitude spectrum of a selected part of the reference incoming signal *in*(3500:4000) is presented in solid line along with its respective reconstructed CELP spectrum (dashed line) and the proposed method's result (dotted line). The spectrum is presented for the frequency range between 0 and 1200 Hz, while all signals are normalised. Note that both techniques attempt to model the spectrum of the signal as accurately as possible. For this particular case the third peak of the spectrum is better modeled by the new method than CELP.
In Figure 4, the same quantities are depicted for a part of the second test incoming signal (Greek sentence) *in*(8200:8700), where the magnitude spectrum is presented in solid line along with its respective reconstructed spectrum at 2400 bps (dashed line) and the reconstructed spectrum at 1200 bps (dotted line), for the normalized signals. In (a) the Frequency range between 0 and 700 Hz is depicted while in (b) between 1300 and 2100 Hz.
Note that despite the similar behavior of the new method at low frequencies for both rates 2400 bps and 1200 bps, the high frequency content at 1200 bps is rather poor. This also explains the drop of the quality of the reconstructed signal at 1200 bps compared to that at 2400 bps.
A new perceptual evaluation procedure had been adopted using the same audience as in the previous experiment. They were asked to indicate whether the sentences listened to, were intelligible or not. At this stage, no need for marking of the best result in terms of quality is required, since it is pretty obvious that the quality of reproduction is analogous to the bit rate used. In Table 3, the subjects' answers are depicted. The majority of the evaluators (80%) indicated that both techniques provide intelligible reconstruction.

Audience	CELP Method's Intelli-gibility	NEW methods' Intelli-gibility	Best Quality
Member_1	YES	YES	CELP
Member_2	YES	YES	CELP
Member_3	YES	YES	CELP
Member_4	YES	YES	CELP
Member_5	YES	YES	CELP
Member_6	YES	YES	CELP
Member_7	YES	YES	CELP
Member_8	YES	YES	NEW
Member_9	YES	YES	NEW
Member_10	YES	YES	CELP

Table 1. Intelligibility Evaluation for CELP standard and the new system at 4800 bps

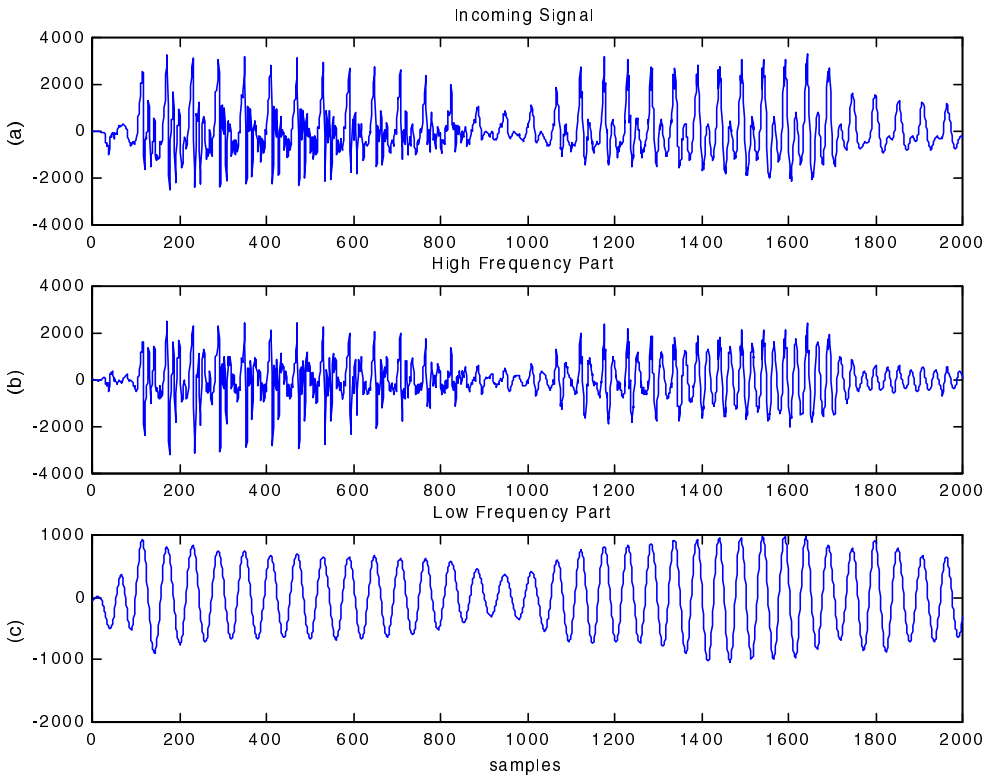

Figure 2: (a) The time domain representation, (b) High Frequency (HF) part and (c) Low Frequency (LF) part of test phrase of the first 2000 samples of "Paris, Bordeaux, Le Mans, Saint-Leu, Léon, Loudun".

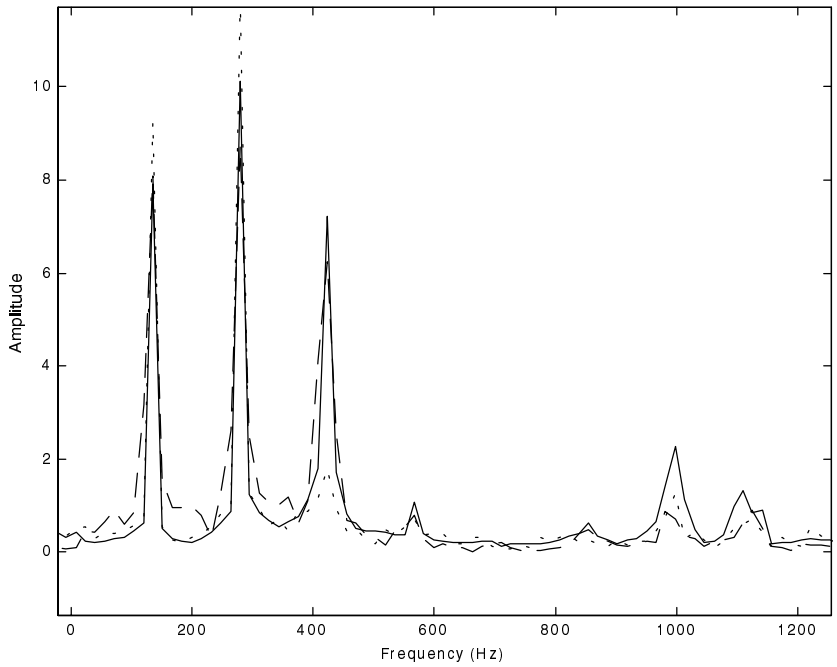

Figure 3: The magnitude spectrum of a selected part of the incoming signal *in*(3500:4000) (solid line) along with its respective reconstructed CELP spectrum (dashed line) and the proposed method's result (dotted line).

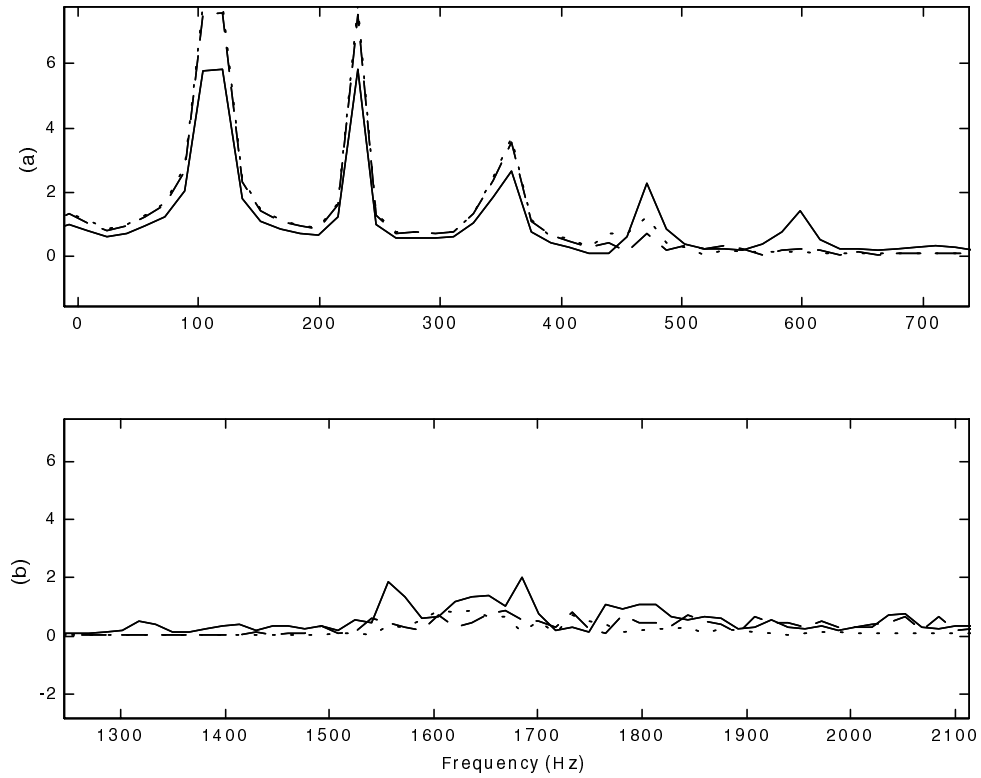

Figure 4: The magnitude spectrum of a selected part of the second test incoming signal *in*(8200:8700) (solid line) along with its respective reconstructed spectrum at 2400 bps (dashed line) and 1200 bps (dotted line).
Frequency ranges: (a) between 0 and 700 Hz and (b) between 1300 and 2100 Hz.

LF part		**HF part**			
Number of Sinusoids	frame length	CodeBook size	CodeBook dimension	Achieved Bit Rate	Segmental SNR
Nsin	lfr	w	N		
3	500	9	70	2400 bps	7.5 dB
3	1000	9	200	1200 bps	3.9 dB

Table 2. Parameter Choice and their corresponding Segmental SNR's and approximate Bit Rates for the new method.

Audience	2400 bps	1200 bps
Member_1	YES	YES
Member_2	NO	NO
Member_3	YES	YES
Member_4	YES	YES
Member_5	YES	YES
Member_6	NO	NO
Member_7	YES	YES
Member_9	YES	YES
Member_9	YES	YES
Member_10	YES	YES

Table 3. Intelligibility Evaluation for the new system at variable bit rates.

Conclusion

In this paper we have presented a new low rate speech coding algorithm, based on a subband approach.

A frame of the incoming signal is first fed to a low pass filter, thus yielding the low frequent (LF) part, while a non-smoothed high frequency (HF) part results by subtraction of the LF from the incoming signal. The HF part is modeled using waveform vector quantisation (VQ), while the LF one via sinusoidal modeling based on SVD (CSE method). At the receiver side, an adaptive postfiltering based on SVD is performed to the HF part and a CSE resynthesis is performed to the LF part, before the two components are added in order to produce the reconstructed signal.

With the appropriate selection of the frame size for CSE and the dimension for the code book, the proposed algorithm results in progressive speech compression and makes therefore possible a variable bit rate scheme. The new method compared to the CELP algorithm at 4800 bps, was proven of similar quality in terms of segmental SNR. Moreover, perceptual evaluation tests proved the new method equally intelligible, while the majority of the evaluators preferred CELP quality. Further perceptual evaluation tests were conducted for the new method at different bit rates and the majority of the evaluators proved it intelligible from this point downward to 1200 bps.

Further research will consist of fine tuning the different parameters involved, namely cut-off frequency, CSE based coding (frame length, number of exponentials), size and dimensions of the code book, aiming at enhancing the speech quality for various bit rates.

References

Campbell, Joseph P.Jr., Thomas E. Tremain and Vanoy C. Welch (1991). "The Federal Standard 1016 4800 bps CELP Voice Coder", Digital Signal Processing, Academic Press, Vol. 1, No.3, pp.145-155.

Doclo, S., Dologlou, I. And Moonen, M. (1998). A novel iterative signal enhancement algorithm for noise reduction in speech. In Proceedings of the ICSLP 1998.

Dendrinos, M., Bakamidis, S. and Carayannis, G. (1991). Speech enhancement from noise: A regenerative approach. Speech Communication, vol.10, no.2, pp.45-47.

Fotinea, S-E., Dologlou, I., Hatzigeorgiu, N. and Carayannis, G. (2000). Spectral Estimation based on the eigenanalysis of companion-like matrices. Accepted for for ICASSP2000, June 5-9, Istanbul, Turkey.

Gersho, A. and Gray, R. (1992). Vector Quantization and Signal Compression, Kluwer Academic Publishers, Boston.

Lemmerling, P., Dologlou, I. and Van Huffel, S. (1998). Variable rate speech compression based on exact modeling and waveform vector Quantization. In Proceedings of Signal Processing Symposium - SPS 98. Leuven, Belgium.