PERCEPTUALLY BASED AND EMBEDDED WIDEBAND CELP CODING OF SPEECH

Alexis Bernard and Abeer Alwan
University of California, Los Angeles (UCLA)
Box 951594, 405 Hilgard Ave, Los Angeles, CA 90095-1594
abernard@icsl.ucla.edu, alwan@icsl.ucla.edu
http://www.icsl.ucla.edu/~spapl

ABSTRACT

This paper presents a novel multi-band CELP coder with the following characteristics: wideband coding (6.5 kHz), variable bit rate (VBR) coding (10-24 kbps), low-delay (10 ms), embeddibility, and perceptually based dynamic bit allocation. The excitation signal of the linear prediction filter is the vector sum of eight off-line pre-filtered bandpass excitation vectors. The eight excitation codebooks are tree structured, providing embeddibility and variable bit rate. The dynamic allocation of the bitstream among the different bands is based on the perceptual importance of each band. The multi-band and perceptual structure of the coding scheme results in graceful degradation with decreasing bit rates both in quiet and in the presence of background noise.

Keywords: wideband, CELP, perceptual, variable bit rate, and embedded coding.

1. INTRODUCTION

The interest in using wideband (50-7000 Hz) speech coding has grown within the last few years since an increased bandwidth can provide more natural and intelligible speech.

New applications for wideband speech arise in the domain of mobile communication, ISDN wideband telephony and videoconferencing. Introduction of a wideband mode is currently being discussed for the forthcoming AMR (Adaptive Multi Rate) codec standard, which will replace the existing GSM codec for its third generation implementation [1].

A wideband speech compression standard was released in 1988 by the CCITT. The G.722 [2] subband ADPCM scheme operates at bitrates of 48, 56 and 64 kbps.

Acknowledgment: This work was supported in part by the National Science Foundation (NSF).

The ITU-T study group 16, pursuing a new standard for wideband coding, specified objectives that include a maximum algorithmic delay of 20 ms, a frame size of 10 ms, and performances at 16 and 24 kbps equivalent to that of G.722 at 48 and 56 kbps, respectively [3]. Traditional approaches to wideband coding include fullband CELP coding [4], and perceptual transform coding [5].

Full band CELP speech coding typically fails to match the spectrum well in the high frequency region, suffers from a high frequency hiss [6].

Transform coders achieve high compression by the use of high frequency-resolution analysis of the signal, requiring a large number of samples per frame, hence a large algorithmic delay that is usually greater than 32 ms.

The approach presented here, perceptually based and embedded multi-band CELP coding, combines high compression ratio of linear prediction with perceptual coding capabilities of a transform coder in the following manner: 1) the separation of the excitation signal into the sum of nearly orthogonal bandpass filtered excitation signals allows the high frequency regions to be represented independently of the other regions, 2) the relative perceptual importance of the bands is derived from a short time signal (10 ms). The coder avoids large delay, is free from hiss at high frequencies and shapes quantization noise according to a perceptual speech metric.

The rest of the paper is organized as follows. Section 2 provides an overview of the perceptual MB-CELP codec. Codebook design operations, codevector search and the perceptual dynamic bit allocation are described in Section 3. In Section 4, we show how different operating bit rates can be obtained. Simulation results and speech quality evaluations are given in Section 5.

2. STRUCTURE OF THE CODER

As mentioned earlier, CELP coding techniques are well suited for low-delay coding and high compression ratios by exploiting short-time domain correlation using linear prediction. As shown in Figure 1, the short-time prediction (STP) filter excitation consists of the vector sum of: 1) a vector chosen from an adaptive codebook in order to remove long-term correlation, 2) eight vectors selected from fixed codebooks. The eight fixed codebooks consist of off-line bandpass pre-filtered Gaussian vectors. The adaptive codebook is composed of the previous excitation vectors. Table 1 shows the frequency region of the different bands. Band selection is the result of informal listening tests. As with bark bands, bandwidths increase with frequency, to reflect the frequency selectivity of the human auditory system.

In our specific implementation, each frame is 10 ms long and is divided into 2 sub-frames of 5 ms each. For each frame, a psycho-acoustic model similar to MPEG [7] computes the masking spectrum of the signal. The masking spectrum is composed of tone-masking-noise and noise-masking-tone components. The resulting global masking spectrum is obtained by taking the logarithmic sum of the masking constituents. The signal to mask (SMR) spectrum is then obtained by subtracting the masking spectrum from the signal spectrum. Figure 2 illustrates the speech spectrum, the masking spectrum for a given frame, and the hearing threshold. The SMR spectrum of the signal is then used to dynamically allocate bits among the different bands and to quantize the line spectral frequencies (LSF) with respect to their perceptual importance.

The linear prediction coefficients are computed every frame. The prediction order *p=16* was found to be reasonable for a 6.5 kHz bandwidth speech signal. The line spectrum frequencies (LSF) are encoded using the switched interframe vector prediction [8] (SIVP) with 4 predictors. The resulting error vector is split vector quantized. The vector is divided into 4 sub-vectors of dimension 4 each and each sub-vector is vector quantized with 6 bits. LSFs are then represented by 26 bits, i.e. 1.675 bit/LSF/frame. In the split VQ codevector searches, the norm of the error vector, which we attempt to minimize, takes into account the distance between LSFs [9], the signal to mask ratio spectrum, and the decrease in frequency resolution for higher frequencies.

The original speech signal and the STP filtered excitation signal are weighted with a perceptual filter of the type $W(z) = A(z/\gamma_1)/A(z/\gamma_2)$ where *A(z)* is the STP filter. The parameters γ_1 and γ_2 are functions of the spectral shape of the input signal, analyzed from a second order linear prediction filter which is turn is obtained as a by-product of the Levinson-Durbin recursion determining the LPC coefficients. If the spectrum is considered flat, then there is less motivation to enhance the valleys with respect to the peaks in the search for vectors, and the parameters are set to γ_1= 0.94 and γ_2= 0.6. If the spectrum is tilted, usually for voiced segments, γ_1 is set to γ_1= 0.98 and γ_2 is adapted to the strength of the resonances, but is bounded between 0.4 and 0.7.

The codebooks' indices, the codebooks' gains, the LSFs, and the bit allocation are transmitted to the decoder (Figure 3). The codebook indices and gains are decoded and the excitation vector is reconstructed. Linear prediction coefficients are also reconstructed from the line spectrum frequencies. The excitation signal is then shaped by the STP filter. Finally, adaptive post-processing operations that enhance speech quality (short-term filter and tilt compensation filter) are performed.

Table 1: Band description in MB-CELP

Band	Band coverage	Bandwidth	# Bark bands
1	[100, 510]	390 Hz	4
2	[510, 1080]	570 Hz	4
3	[1080, 1720]	640 Hz	3
4	[1720, 2320]	600 Hz	2
5	[2320, 3150]	830 Hz	2
6	[3150, 4100]	950 Hz	1.5
7	[4100, 5300]	1200 Hz	1.5
8	[5300, 6500]	1200 Hz	1

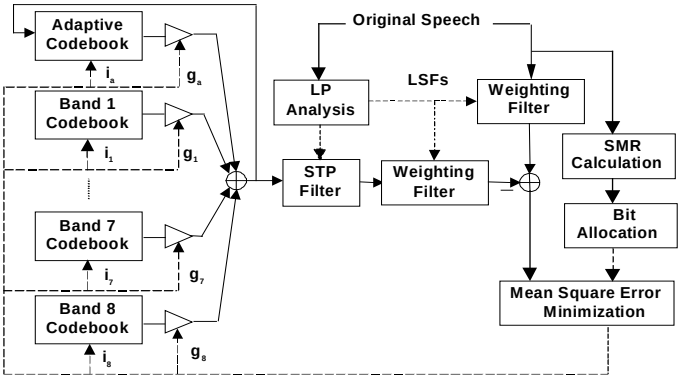

Figure 1 : MB-CELP encoder structure.
Dotted lines reflect the transmitted information.

3. SEARCH OF THE MULTI-BAND CODEVECTORS AND GAINS

For each sub-frame, the long-term predictor vector and its gain are first calculated and the resulting vector is removed from the excitation vector to be searched. With a 5 ms long subframe and a sampling frequency of 13000 samples/s, the optimal integer delay is found in the previous subframes and can be represented with 7 bits. The mean removed logarithmic gain is non-uniform scalar quantized with 4 bits. For the second subframe, the pitch delay and adaptive codevector gains are searched around the values of the first subframe. They are represented using 3 bits only.

The remaining excitation, after pitch prediction removal, is then vector quantized using multi-band shape-gain VQ. The 8 fixed excitation codebooks are pre-filtered codebooks with non-uniform bandwidth, as described in Table 1. It must be noted that since the filtered codevectors are of finite duration (equal to the subframe length), there is leakage between adjacent codebooks. However, since the codevectors of the different codebooks are nearly orthogonal, the sequential search of the codebooks provides almost the same performance as that of an optimal joint search, but with a greatly reduced complexity. Furthermore, being sequential, spectral leaking of the i^{th} band can be compensated for in the choice of the $i+1^{th}$ codevector.

The eight training sets of pre-filtered vectors are clustered into a 3 stage TSVQ with successively 16, 16 and 4 splitting branches. At the first level of the tree, the entire training set is clustered using the k-means algorithm into 16 codevectors and the training set is divided into 16 subsets. The same k-means algorithm is then applied on the sub-sets to find the second level of codevectors and subsets. Such a tree-structure in vector quantization allows for embeddibility and variable rate quantization. For each band, one can allocate 4, 8 or 10 bits. Other types of VBR vector quantizers include multi-stage, entropy and pruned VQ [10].

In the search for the best codevectors, more bits are allocated to the bands that account for the most perceptual content of the signal. Based on the MPEG psycho-acoustic model, the minimum signal to mask ratio (SMR) is derived for each band. The bands are sorted in increasing order of importance. Once the bands are sorted, the codevector search starts with the most significant band, and ends with the least significant one. The actual bit allocation for each band depends then on the target bit rate and its ordering. With 8 bands, there are 40,320 possible ordering combinations. However, the 32 most frequent ordering combinations account for more than 90% of the actual band ordering combinations. Only 5 bits are then used to represent the bit allocation.

The logarithm of the mean removed codevectors gains are split vector quantized. The 8 gains are split in 4 two elements vector and each vector is in turn quantized with 5 bits, for a total of 20 bits.

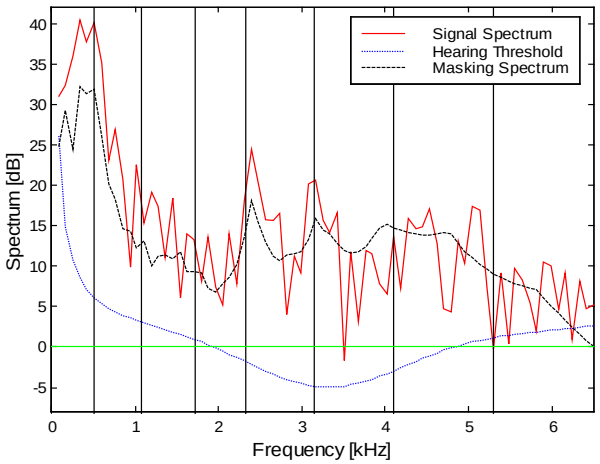

Figure 2: Signal and masking spectra illustration

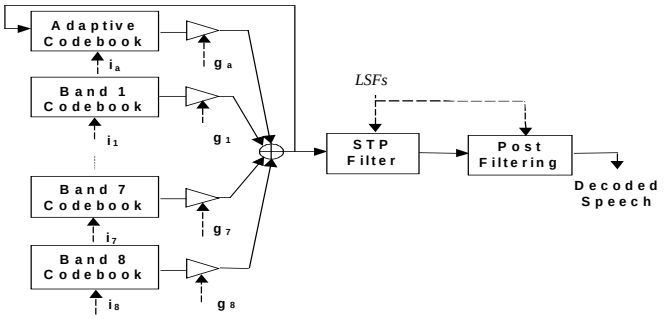

Figure 3: MB-CELP decoder structure

4. OPERATING BIT RATES

Depending on the number of bits used to represent the excitation vectors and their update rate, one can encode speech with total bitrates ranging between 10 and 24 kbps. Lower rate codes are embedded into higher rate codes. In other words, a lower rate encoder can be obtained by simply truncating the bitstream of a higher rate encoder. This also means that bits can be discarded or dropped between the encoder and the decoder. Finally rate variability, embeddibility and bit allocation based on the perceptual characteristics of the signal leads to graceful speech quality degradation with decreasing bitrates.

Table 2 indicates how different rate encoders (10-24 kbps) are obtained by modifying the total number of bits used to represent the fixed codevector indices and by changing their update rate. For low bit rate coders, where the codevector indices and gains are updated only once a frame, the update takes place for the second sub-frame. The excitation codevectors and gains for the first sub-frame are obtained by linear interpolation between adjacent sub-frames.

Table 2: Bit rates operating modes

Rate	Fixed	1st subframe		2nd subframe		Bits/
kbps	bits	index	gains	index	gains	frame
10	48	32	20	0	0	100
12	48	52	20	0	0	120
14	48	72	20	0	0	140
16	48	36	20	36	20	160
18	48	46	20	46	20	180
20	48	56	20	56	20	200
22	48	66	20	66	20	220
24	48	76	20	76	20	240

The 48 fixed bits divide as follows: 5 bits for the bit allocation, (7+3) bits for the long-term predictor delays, (4+3) bits for the long-term predictor gains, and 26 bits for LSFs quantization.

5. RESULTS

An informal listening test indicates that at high bit rates, the new perceptual MB-CELP performs almost transparent coding. For the targeted low bit rates, quantization noise is efficiently shaped after the masking spectrum, resulting in graceful speech quality degradation. Table 3 shows the segmental signal to noise ratios (SEGSNR) obtained while testing the coder in two situations: clean speech and in the presence of speech shaped background noise at 10 dB SNR. Test material consists of 4 sentences (2 male, 2 female) taken from the TIMIT database. Simulation results also indicate graceful degradation.

6. CONCLUSIONS

We have presented a new perceptual and embedded MB-CELP coder. Dynamic bit allocation based on signal to mask ratios of the bands and a tree-structure vector quantization scheme for the fixed excitation codebooks allow the coder to be embedded, variable bit rate, and to degrade gracefully at lower bit rates. Frame size is limited to 10 ms and the algorithmic delay is kept limited, due to the absence of any analysis-synthesis filterbank and the tree structure of the codebooks which limits the maximum code searches to $8 \cdot (16+16+4) = 288$ per frame. Rate variability of the coder, together with its embeddibility make the coder flexible for situations when channel conditions require rapid changes in the source coding bit rate or where bits need to be dropped anywhere in the communication link. The coder complies with the ITU requirements for wideband coding, but does require the storage of 8 codebooks.

Table 3: Segmental SNRs for clean and noisy signals

Bit Rates (kbps)	Clean Speech SEGSNR (dB)	Noisy Speech SEGSNR (dB)
10	7.35	7.08
12	8.45	7.95
14	9.34	8.32
16	9.37	8.36
18	9.64	8.83
20	10.18	9.48
22	10.59	9.97
24	10.96	10.23

7. REFERENCES

[1] ETSI SMG11-AMR, "Draft Adaptive Multi-Rate Study Phase Report", Tdoc, Version 0.4, Aug. 1997

[2] CCITT, "7 kHz Audio Coding within 84 kbit/s", in *Recommendation G.722*, vol. Fascile III.4 of *Blue Book* pp. 269-341, ITU, Melbourne, 1988

[3] ITU-T SG16 Q.20, "Reference for the ITU-T wide-band (7 kHz) Speech Coding Algorithm", Apr. 1997

[4] E. Harborg, J.Knudsen, A. Fuldseth, F. Johansen, "A Real-Time Wideband CELP coder for a Videophone Application", Proc. ICASSP 1994, pp. II 121-124

[5] J. Johnston, "Transform coding of audio signals using perceptual noise criteria", IEEE J. Selected Area in Communications, 6:314-232, 1988

[6] A. Ubale, A. Gersho, "A Low-Delay Wideband Speech Coder at 24 kbps", Proc. ICASSP, Seattle, 98.

[7] Brandenburg, G. Stoll, "ISO-MPEG-1 Audio: generic Standard for Coding of High Quality Digital Audio", J. Audio Eng. Soc., Oct. 1994, Vol. 42, pp. 780-792

[8] M. Yong et al. "Encoding LPC spectral parameters using switched adaptive inter-frame vector prediction", Proc. of ICASSP 1988, pp. 402-405.

[9] Laroia, R.; Phamdo, N.; Farvardin, N., "Robust and efficient quantization of speech LSP parameters using structured vector quantizers". ICASSP 91, 5, 641-644

[10] Lookabaugh, T.; Riskin, E.; Chou, P.; Gray, R. "Variable rate vector quantization for speech, image, and video compression" IEEE Trans. Com., vol.41, Jan. 1993. pp. 186-199.