Varible Low Bit Rate CELP Speech Coder

Şef lucrări Cornel Balint
Politehnica University Timisoara

ABSTRACT. The paper presents a variable low bit rate CELP (Code Excited Linear Prediction) speech coder. The coder combine three different coding structures selected in accordance with short time speech characteristics. In order to select a adequate coder structure, the input speech is classified in three categories: non-speech, non-voiced speech and voiced speech. A speech-non-speech classification method using spectrum envelope variation vas developed in orders to reduce the influence of environmental noise. For each category, an appropriate coding method was selected. The proposed coder is a classical CELP coder, adapted to the input speech at every speech frame. The average bit rate is approximately 2,5 kbps, depending of the activity of speech source.

1 Introduction

The variable rate speech coder vas developed in order to reduce the bit rate according of nature of speech. The general structure of proposed coder is in accordance with a CELP coder. The speech input signal is organized in frames. For each frame, a signal analysis is performed, in order to calculate the LPC coefficients and the residual speech signal, that is coded using a codebook in a synthesis by analysis procedure. The CELP coder obtains a bit rates lower as 16 Kbit/sec, acording of speech quality. For narrow band speech the CELP coder obtains a bit rate lower as 4,8 Kbit/sec, preserving the speech naturally and speaker recognizability. In order to obtain a lower bit rate, each frame of input speech signal is classified in speech or non-speech frame and different coding methods are using for speech and non-speech frames.

2 Speech/Non-Speech Classification

For the noise (non-speech) frame, the coding method is generally designed for a very low bit rate, usually under 1 kbps, in opposite with speech frame when a large bit rate is necessary. In this case, major quality degradation occurs when a speech frame is classified and coded as a noise frame.

Signal classification in speech or noise can use a variety of indicators. The frame power can be considered as a good indicator, but in noise presence the frame total power becomes inefficient.

The prediction gain is an other indicator with the same inconvenient.

The proposed coder use for signal classification the spectrum envelopes modification. The measure of spectrum modification and power variation is used in each frame for the decision non-speech/speech. For each signal frame, the spectrum envelope (SE) modification is computed using the prediction parameters a_i for the current frame and a average set a_{avi} of prediction parameters computed for non-speech frames. The averages set of prediction coefficients and total power for non-speech frames are updated in each frame by the current value if the actually frame is classified as non-speech.

The spectrum envelope (SE) modification is defined by:

$$SE = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}\left(\log_{10}\frac{\left|1-\sum_i a_{avi}\exp(2\pi jni/N)\right|^2}{\left|1-\sum_i a_i\exp(2\pi jni/N)\right|^2}\right)} \qquad (1)$$

The total power of current frame P and power variation PE is also computed by subtracting the average power of non-speech frame P_{av} from total power of actual frame:

$$PE = P - \alpha P_{av} \qquad (2)$$

were α is a constant coefficient.

The classification algorithm takes the decision according to rules:

If PE>0, the frame is classified as speech frame,

Else:

If SE>SET, the frame is classified as speech frame,

Else: the frame is classified as non-speech frame.

with SET a threshold depending of the average power of non-speech frames.

After a frame classification, α , P_{av} and SET are updated according of the new frame and category of precedent frame:

$$P_{av}new = (1-\lambda)P - \lambda P_{av} \quad (3)$$

3 The Coder

The bloc diagram of the proposed coder is presented in figure 1.

The input signal is organized in frame, each frame of 30 ms or 240 samples, at 8 KHz sampling frequency, having 4 subframes of 7,5 ms or 60 samples.

For each frame, a LPC analysis are performed. If the current frame is a speech frame, 10 LPC parameters are used. If the current frame is a non-speech frame, only 4 LPC coefficient are used in a simplified way, in order to obtain a bitrate reduction.

The LPC parameters are cuantized in line spectral pair (LSP) representation, a method that provide efficient cuantization.

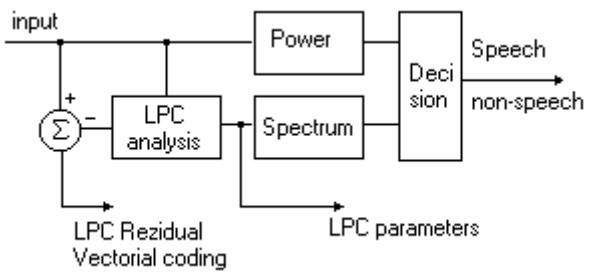

Fig. 1. Coder bloc diagram

LPC parameters have no bound and so is difficult to define the quantization region and small quantization error may cause unstable syntheses filter. In opposite with LPC, line spectrum pair LSP are resonant natural frequency of vocal tract and are bounded naturally by human, psychical features. LSP are also naturally ordered and the order of parameters must not to be transmitted. Because LSP represent the resonant frequency of human vocal tract, they will not change radically during a speech frame. The reduction of bit rate is possible by exploiting this property, using only a set of LSP during a frame and interpolate the LSP for each subframe, according to relations:

$$LSP_i^1 = (7/8)f_i + (1/8)g_i \quad , \quad LSP_i^3 = (3/8)f_i + (5/8)g_i$$
$$LSP_i^2 = (5/8)f_i + (3/8)g_i \quad , \quad LSP_i^4 = (1/8)f_i + (7/8)g_i \qquad (4)$$

with the superior index denoting the subframe, as in figure 2.

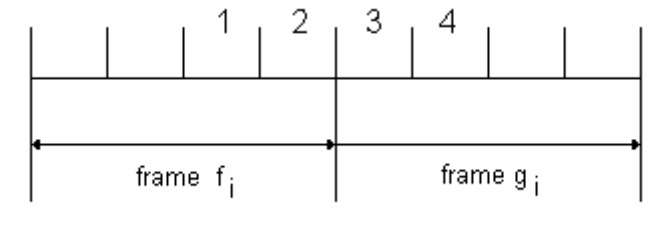

Fig. 2. Subframe interpolation of LSP

The speech prediction residual is aplied to a vectorial coding scheme, using two codebook in order to generate a correct representation of residual for the two categories of speech signal: speech or non- speech.

The excitation for the synthesis filter is generated by two codebook: a adaptive codebook and a stochastic codebook.

For speech unvoiced signal, the adaptive codebook is eliminated, and the excitation signal is obtained from stochastic codebook only. An algebraic codebook structure with ternary vector (that use only –1, 0 and +1 value) is proposed, in order to eliminate a large number of multiplication.

For the non-speech signal, the excitation for the synthesis filter is a withe noise locally generated at reception, in order to reduce the bit rate and only the gain is transmitted.

A complexity reduction is proposed using a relative quantization of pitch. The pitch search is performed in a relative way exploiting not only the correlation of the pitch in consecutive subframe, but also the corelation of pitch over consecutive frame. A different search domain for pitch is used in each subframe of the current frame and in the next frame.

	Non-speech	Speech
LPC	9	22
Adaptive codebook	-	15
Stochastic codebook	-	9x4
Gain	5	5x4
Mode	1	1
Bit/frame	15	94
Bit rate	15x1000/30=500 bit/sec	94x1000/30=3,13 Kbit/sec

Depending of speech signal, the achieved average bitrate is lower as 2,4 Kbit/sec.

4 Voice Activity Detector (VAD)

The nature of human speech determines the discontinuity of speech signal. In that times intervals, the information carried by the signal is irrelevant and the signal is not to be coded or transmitted, resulting a supplementary bitrate reduction. In order to keep the naturally speech, the speech signal is transmitted even in that times period, but the coding scheme use a minimum bitrate.

The decision of activity of speech source is made by the voice activity detector (VAD). A god measure of voice activity is the total energy of analyzed frame or subframe and the energy tresholding method can work satisfactorily in high signal to noise ratio. Considering the situation with a low signal to noise ratio and the varying nature of ambient noise, the energy tresholding method is completed by additional steps.

The energy tresholding method use the total power of current frame, calculate in order to determine the nature of frame, and a energy treshold B, computed as a function of background noise level, updated each frame:

$$B = \min(R(0), \max(B_{prev}, B_{prev+1})), \quad (5)$$

were R(0) is the autocorrelation value used in LPC algorithm and B_{prev}, B_{prev+1} are the background treshold for the previous frames.

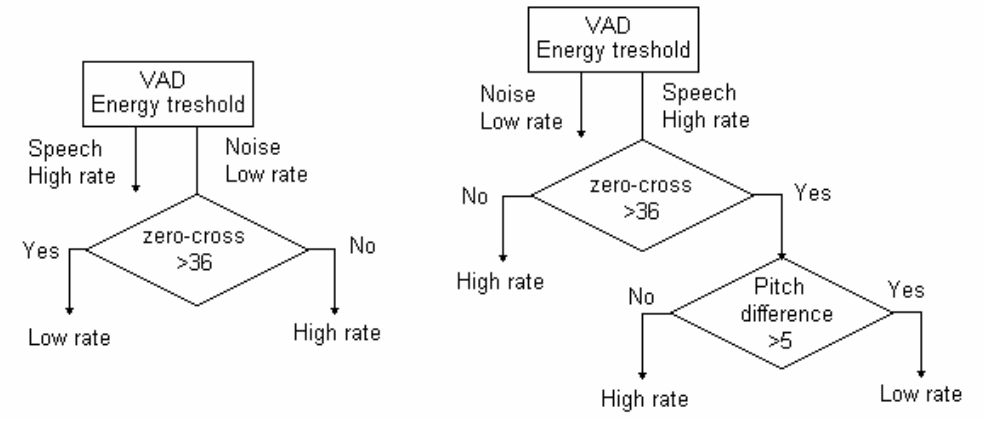

Fig. 3. Bit rate decision algorithm

In the second step, the energy tresholding is confirmed by the number of zero-crossing, as a measure of presence of lower frequency component in speech signal and by the difference of pitch lag in the current and previous subframe. For the voiced segment of speech signal, that difference must be around zero. The complete decision algorithm is illustrated in figure 3.

References

[AS85] **B.S. Atal, M.R.Schroder**, *Code Excited Linear Prediction (CELP)*, IEEE Proc. on ASSP, Tampa, Florida, 1985

[CK99] **W Chung, S. Kang**, D*esign of a Variable bit rate Algorithm for the CS-ACELP Coder*, IEEE Trans. on Inf. and Syst., vol E82, no.10, 1999

[ETSI99] **ETSI standard EN 300 973**, *Digital Cellular Telecommunications System, Half Rate Speech, Voice Activity Detector*, 1999

[GG93] **A. Gersho, R. M. Gray**, *Vector quantization and signal compresion*, Kulwer Academic Publications, Boston 1993

[GP92] **A. Gersho, E. Paksoy**, *An Overview of Variable Rate Speech Coder for Cellular Networks*, IEEE Proc. ICWC, 1992

[LBG80] **Y. Linde, A. Buzo, R.M. Gray**, *An Agorithm for Vector Quantizer design*, IEEE Trans. On. Communications, vol.28, 1980

[OA98] **M. Oshikiri, M. Akamine**, *A 2,4 kbps Variable rate ADP-CELP Speech Coder*, ICASSP98, Seatle, Washington, 12-15 may 1998