Pitch Controlled Variable Bitrate CELP Speech Coding

Robert M. Oberhofer and Frank J. Owens

Dept. of Mech. & Electr. Engineering, University of Ulster, Shore Road, Jordanstown, BT37 0QB, Northern Ireland

Abstract: In the past decade Code Excited Linear Prediction (CELP) has become the dominant speech coding algorithm for bit rates between 4 kbps and 16 kbps. Traditionally, CELP type algorithms produce a fixed bit-rate. However, a fixed bit-rate is not a requirement anymore for modern telecommunication and computer networks as well as for voice storage applications. The variable bitrate CELP coder presented in this paper features, besides the usual classification and bit allocation schemes, additionally a variable analysis frame size, which is adapted according to the pitch of the speech signal.

INTRODUCTION

The standard 1016 CELP speech coder [1] is perceived as producing good quality speech at 4800 bps. However high pitched speech from typically female or children speakers is often reproduced with some 'roughness'. There are therefore two main points for the motivation to implement a pitch controlled variable bitrate speech coder:

1. A reduction of the overall bit-rate of the standard 4800kbps CELP speech coder by implementing a variable bit-rate scheme. This is implemented through common classification techniques.

2. Addressing the 'roughness' of the coder for higher pitched speakers through adapting the analysis frame size according to the pitch of the speaker. Higher pitched speech is analysed in shorter intervals to increase the resolution and therefore quality.

The codec described in this paper belongs therefore to the class of source-controlled variable rate codecs.

SYSTEM OVERVIEW

The speech coder consists of three main parts as shown in the block diagram. The Voice Activity Detection (VAD), The Frame Extractor and Classifier and the CELP type speech coder.

**FIGURE 1**: Block diagram of codec

*Voice Activity Detection*

Voice Activity Detection is a very effective means of reducing the bit rate of a speech coder. A simple method has been implemented using energy computation and zero crossing rates compared against threshold levels. Voice inactivity is indicated by a flag and decoded with silence.

*Frame Extractor and Classifier*

The heart of the coder intends to extract features out of the signal and to classify them. Typical classes of speech are voiced, unvoiced, transitional and stationary.

The pitch period is estimated by standard auto-correlation techniques. However the auto-correlation is just performed at lags, where the speech signal shows peaks as in [2]. The distance between adjacent
high peaks correlates to the pitch in voiced speech. This reduces significantly the necessary computation while producing good results for the classification. The analysis frame size is set accordingly to either 45, 55, 65 or 75 samples (8000Hz sampling frequency). Unvoiced speech does not have a significant pitch period, and this type of speech is therefore analysed using a constant size of 60 samples.

Onsets of speech are often characterised by a rapid change of the speech situation. To accommodate this, the analysis frame size is set to initially 45 samples to allow a swifter adaptation of the coder to the new situation.

The transitional / stationary criterion is used to allow for a more effective coding. Static speech shows no significant change of speech parameters between frames. The parameters can therefore be coded differentially.

**CELP Coder**

The CELP module performs the main task of coding the speech. It uses the CELP components of Linear Prediction Analysis, Pitch Prediction and Excitation Codebook Search in a variable bitrate environment.

The Linear Prediction Analysis was initially performed on the determined framesize of the Frame Extractor and Classifier (45, 55, ..., samples). This proved to be not just inefficient but also unnecessary. The Linear Prediction is therefore performed, as in the original CELP coder, every 240 samples (30ms). The coefficients are transformed into Line Spectral Pair (LSP) parameters. The LSPs are transparently quantized with a split vector quantisation scheme at 24 bits per frame.

Since the Linear Prediction Analysis frame is fixed, and the analysis frame for Pitch Prediction and Excitation is flexible, these two run concurrently but independently as shown in Figure 2:

<table>
<thead>
<tr>
<th>Linear Prediction Frames</th>
<th>Analysis Frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>240</td>
<td>240</td>
</tr>
<tr>
<td>65 65 45 55 65 65 65 55 60 60</td>
<td></td>
</tr>
</tbody>
</table>

**FIGURE 2: Overlapping of Linear Prediction frames and Pitch Prediction and Excitation.**

The Linear Prediction parameters are interpolated for each individual analysis frame.

The Pitch Prediction is performed as in the original CELP coder for voiced frames. The pitch codebook is searched exhaustively for non-static speech and a subset thereof is searched for static speech. The pitch is coded differentially for static speech. Speech frames marked as unvoiced are not coded by the Pitch Predictor.

The Excitation Codebook is searched exhaustively in all situations. The codebook is the stochastic ternary codebook of the original 1016 CELP coder [1].

**CONCLUSION**

The presented speech coder is capable of producing CELP quality speech at a bit rate of 3200 to 4000 bps depending on the speech situation. For clean, noise-free speech, this figure can be reduced to typically 2600 - 3200 bps with the Voice Activity Detection included. The Voice Activity Detector is however less efficient in noisy speech due to its simplicity. A more sophisticated implementation is therefore desirable.

The average computational load is lower than that of the original CELP coder. However the peak computational effort is above CELP due to the additional overhead and possible shorter frame sizes. This makes the algorithm less economic for typical telephony applications, but more efficient for applications where the speech can be sufficiently buffered or processed offline, such as digital speech broadcasting, or voice storage applications.
