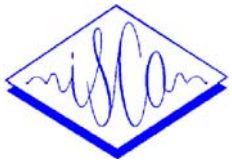

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

SOURCE-DEPENDENT VARIABLE RATE SPEECH CODING BELOW 3 KBPS

M. Stefanovic, A. Kondoz

Centre for Communication Systems Research (CCSR)
University of Surrey, Guildford, Surrey, GU2 5XH, United Kingdom
M.Stefanovic@ee.surrey.ac.uk, A.Kondoz@ee.surrey.ac.uk

10.21437/Eurospeech.1999-338

ABSTRACT

This work addresses the need for high quality, very low bit rate speech compression algorithms that can be utilised in many forthcoming multimedia applications. The speech coding algorithm proposed in this paper is a variable rate system based on the adaptive source-driven frame length scheme. Maximum speech compression is achieved for long-term steady-state speech and non-speech (silence and unvoiced) conditions. In addition, shorter frame sizes are used to code those difficult-to-model speech transitions, thus improving the overall perceptual quality when compared with traditional fixed rate schemes. This codec may be used for implementing various voice communication systems, such as Voice Store and Forward systems and digital answering machines, or to augment low bit rate integrated digital packet networks.

1. INTRODUCTION

Advances in very low bit rate speech coding at rates below 3 kbps are reaching saturation in terms of achieving their main goal - lowering the bit rate, whilst preserving relatively high perceptual quality of speech. A major part of research effort has been mainly concentrated on improving the well-known speech coding methods, such as CELP codec [1] and MBE vocoder [2]. The common feature of these codecs is that they operate based on fixed frame sizes in the range of typically 20-30 ms. However, as speech is generally considered non-stationary and bursty in character with short term statistics that vary substantially with time, fixed frame size structure restricts the coding process from being truly optimised.

Recent research [3][4] has shown that variable rate techniques can augment fixed rate speech coding systems, producing similar or higher quality speech at lower average frame rates. However, these variable rate coding techniques still preserve the fixed frame length structure. A more sophisticated approach to variable rate coding is to adapt the encoding frame length according to the variation in the short term speech characteristics. Thus, parameter transmission occurs more frequently when speech characteristics are changing rapidly, as in phoneme transitions, and less frequently when speech characteristics are relatively constant as in steady-state sounds. In this way, not only is there a potential for improving the perceptual speech quality, but also for reducing the average frame rate.

This paper focuses on the novel variable frame length segmentation method, which has been applied to fixed coding structure of the Split Band (SB) LPC vocoder operating at 2.6 kbps. The optimum size of the segmented frames is found according to the four-level source-dependent segmentation algorithm and can vary between 10 ms and 40 ms. The subjective test results have shown that the variable frame length vocoder can improve its fixed frame length performance, especially during speech transition regions. Moreover, the vocoder has shown significant reduction in the average frame rate of around 22% for voiced speech.

2. SPEECH SEGMENTATION

The output speech quality of fixed frame length codecs largely depends on their ability to code the most perceptually difficult regions of speech, such as dynamically changing speech transitions. The fixed frame length algorithms typically attempt to model these regions by relying on the complex interpolation schemes, which can only provide a partial solution. However, a more practical approach is to make the frame length adaptable. In the case of speech transitions, there is a clear need to reduce the analysis frame size. By shaping a transition with a sequence of shorter frames, one can expect to be more successful in detecting and modelling its rapid dynamics, thus improving the overall coding quality. Shorter frames rely less on the use of sophisticated interpolation techniques, except for eliminating the block edge effects. The obvious drawback though is an increase in frame update rate.

Successful application of the shorter frame sizes inevitably calls for the use of longer frame sizes for speech regions that do not significantly change their characteristics (e.g. pitch period, voicing and LP shape) over a long period of time, such as silence, unvoiced and steady-state voiced with quasi periodic features. However, the problem of potential averaging of the

short-term energy within a longer analysis frame must be solved by making use of modelling parameters, such as either time envelope or spectral amplitudes, which are able to adequately reflect its changing properties. The biggest advantage of utilising the longer frame sizes is the reduction in the overall frame rate. However, the perceptual quality already achieved by the fixed frame length scheme for the steady state regions must not be compromised.

The most important part of the variable frame length system is the accurate and robust speech segmentation algorithm with source-dependent criteria. The segmentation criteria are required to produce a speech coder whose overall output quality is preferably better or at least matches that of the equivalent fixed frame length system. As it is a source-dependent algorithm, the average frame rate is expected to constantly vary; a speech material with predominantly short bursts of speech will require a higher update rate, whereas the one that contains longer steady-state speech periods will have its update rate significantly reduced.

3. CODING STRUCTURE

The underlying SB LPC vocoding structure, which supports the variable frame length scheme, is initially set to operate at 2.6 kbps and has shown to produce high quality synthetic speech [4]. The fixed rate vocoder processes 20 ms frames of DC rejected, high frequency speech. Its bit allocation scheme is given in Table 1 below.

Table 1. Bit Allocation Scheme

Parameter	**Number of bits per 20 ms frame**
Pitch	7
Voicing Decision	3
LSF	24
Energy	12
Shape	6
Total	52

3.1 ENCODER

Figure 2 shows the block diagram of the encoder. Speech analysis is performed on speech frames in the range from 10 ms (80 samples) to 40 ms (320 samples) depending on the parameter characteristics of the 8 kHz-sampled speech. The detection of the pitch period is performed by the Sinusoidal Model Matching (SMM) based pitch detection algorithm [4] and the Pitch Refinement Function [4] operating in the frequency domain. A split-band voicing decision [4] is also made in the frequency domain once per frame. The tenth order LP filter is used for spectral modelling. The LP coefficients are quantised in the LSF domain using a 24-bit split vector quantiser based on moving average prediction. The quantised LP parameters are then used to filter the LP residual, which is required for determining the excitation harmonic magnitudes. The frequency domain amplitude calculation is performed once per half frame and the windowing of the half frame sized LP residual is pitch period dependent.

Estimating harmonic magnitudes twice per frame improves the accuracy of the amplitude determination, especially for longer frame sizes where pitch and voicing may stay relatively constant, but the amplitude content can vary more significantly. In the case of voiced harmonics, the amplitude calculation uses the Minimum Squared Error matching method applied in the frequency domain. In the case of unvoiced harmonics, the amplitude calculation determines the RMS spectral energy over the unvoiced harmonic band. The quantisation of the two sets of amplitudes of the variable frame length requires 6 more bits than the typical one set per frame scenario. The extra bits are needed for the RMS energy of the additional set. The bit allocation for quantising the shape remains unchanged.

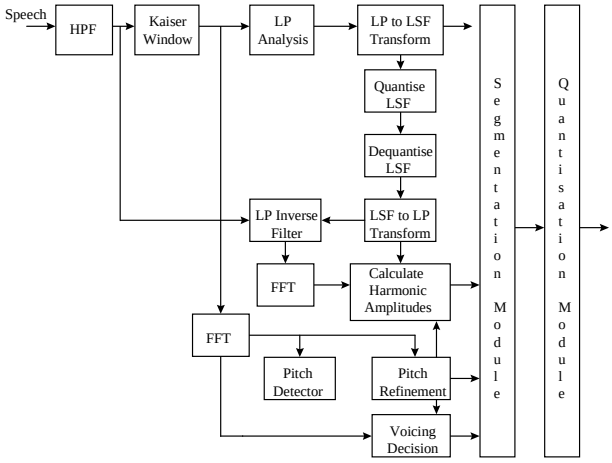

Figure 2. Encoder Block Diagram

The most important part of the variable frame length encoder is the accurate and robust speech segmentation module where the search for the most optimal frame size is source-dependent. The 80-sample or 120-sample starting frame, the size which is dependent on the calculated pitch period, is analysed for its parameters and then extended by 40-sample segments up to a maximum length of 320 samples. At each stage of frame extension, the speech parameters are recalculated and compared with those that characterised the starting frame using the novel four-level 'search-and-compare' decision algorithm shown in Table 2. The thresholds for each parameter comparison stage were derived from both the subjective and objective tests.

Table 2. A four-level speech segmentation decision

Decision Level	Parameter	Threshold
1.	Pitch	≤±1 sample
2.	Voiced Harmonics 2-12 13-18 19-24 25+	 ≤±2 ≤±3 ≤±4 ≤±5
3.	Envelope Energy	≤50%
4.	Spectral Comparison All Voiced Harmonics 1st Voiced Harmonic Remain. Voiced Harmonics	 ≥Av. Energy ≤±15% ≤±30%

The *pitch period* threshold of ±1 sample has been derived objectively. It is a coarse measure that usually permits a large volume of frames to be considered for the frame extending process and, thus, represents the first decision level. If the pitch fluctuates by a larger margin, the frame extending process is terminated; otherwise, the frames are passed onto the other speech parameters for a finer threshold selection.

The range thresholds determining the fluctuations in the number of *voiced harmonics* were derived through subjective listening tests. They clearly highlight the importance of the low frequency harmonic structure, as only a minimum fluctuation of ±2 is allowed within the first 12 harmonics.

The *envelope energy* is permitted to vary up to a maximum of 50% of its starting value. This threshold aids the onset indication and prevents the use of longer frames for speech containing large energy fluctuations.

The fourth decision level concerns the spectral behaviour during the frame extending process. The *voiced harmonic magnitudes* of the spectrum of the starting frame are individually compared against the voiced harmonic magnitudes of the spectra of the frame extensions. The fluctuation in the harmonic magnitudes has to meet three different energy thresholds, as detailed in Table 2, before frame extending is continued.

The optimum frame that is encoded is the longest frame whose speech parameters lie within the thresholds at all four decision levels. As a rule of thumb, the speech transitions are usually modelled with shorter frames, whilst the slowly-varying speech segments and the unvoiced regions (including silence) are modelled with longer frames, examples of which are shown in Figure 3.

In order to notify the decoder on the size of the encoded frame, an additional 3-bit header is also incorporated in the bit rate structure yielding a total bit rate of 2.75 kbps.

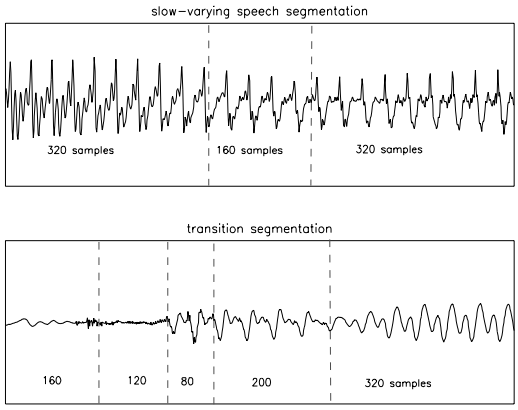

Figure 3. Frame segmentation example for slow-varying and fast-varying speech

3.2 DECODER

Figure 4 represents the decoder schematic of the SB LPC vocoder operating at the variable frame rates. The *LSF parameters* are decoded, interpolated and transformed into the LP coefficients used in the LP synthesis filter. The decoded *voicing* and *pitch period* control the amount of voiced and unvoiced excitation generation. The *amplitudes* are decoded and then modified in such a way as to enhance the perceptual performance of the vocoder. The voiced excitation is generated using the time domain harmonic summation of the voiced amplitudes. The unvoiced excitation is generated in the frequency domain and then transformed into the time domain where it is combined with the voiced excitation. Finally, the combined excitation signal is passed through the LP filter to produce the synthetic speech.

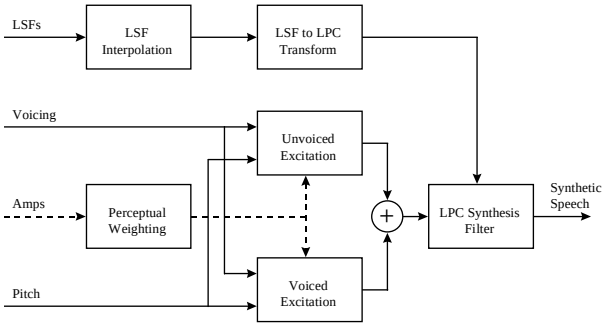

Figure 4. Decoder Block Diagram

The synthesis structure of the variable frame length decoding algorithm is based on a *half frame delay* adaptive interpolation mechanism. The half frame delay mechanism means that the synthesised frame is usually of a different length to the two encoded frames whose parameters are being interpolated at the decoder side, as shown in Figure 5.

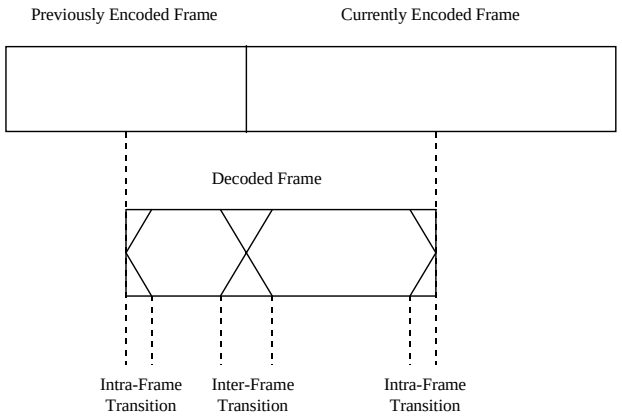

Figure 5. Block diagram of the Adaptive Interpolation Method used

The adaptive interpolation algorithm operates at the intra-frame and inter-frame transition level, as illustrated in Figure 5. The intra-frame transition lies between the two half frames of a given encoded frame. As each half frame uses a separately transmitted set of amplitudes, a transition region of a constant short size is established to smooth out any significant changes, especially during long frames. The variably sized inter-frame transition links the adjacent encoded frames. The selected size of the transition depends solely on the paired combination of the encoded frame lengths. The transition size is set to range from 2 sub-frames to 6 sub-frames. The biggest advantage of this adaptive interpolation method lies in its low complexity and a higher degree of overlap.

4. PERFORMANCE

The performance of the variable frame length structure of the SB LPC vocoder was assessed using the MOS scale, whereby the subjective quality of the variable frame length scheme was compared with the subjective quality of the fixed frame length scheme. The relevant fixed rate speech coding standard (4.15 kbps I-MBE) was also included in the tests as a reference point. Twenty-four subjects, mostly experienced listeners, were used to conduct the informal MOS listening test that included two male and two female sentences.

Table 3. MOS Test Results

Coder	Frame Length (samples)	MOS	Variance	Change in average frame rate
4.15 kbps I-MBE	**160 (fixed)**	**2.93**	**0.29**	**None**
2.6 kbps SB LPC	**160 (fixed)**	**3.54**	**0.52**	**None**
2.75 kbps SB LPC	**80-320**	**3.51**	**0.52**	**+1%**
	120-240	**3.63**	**0.41**	**+2%**
	120-320	**3.47**	**0.50**	**-2%**
	160-320	**3.52**	**0.29**	**-22%**

The test results, as given in Table 3, show that the perceptual quality and the change in the frame rate much depend on the employed frame size range of the variable frame length codec.

The short range version that permits frame extending only (160-320 samples) produced similar quality of speech as the fixed rate codec. However, its real gain over the fixed rate structure is the average drop in voiced frame rate by 22%. The moderate range version (120-240 samples) performed better than the fixed rate codec with an insignificant increase in frame rate. Both the short and the moderate range versions represent the real success of the variable frame length coding.

The full range versions (80-320 samples and 120-320 samples) performed only slightly worse than the fixed rate codec, although they did not experience any large fluctuations in frame rate. This decrease in speech quality indicates the level of difficulty and complexity involved in jointly coding very short and very long frames. However, the tests also confirmed the importance of selectively using highly accurate shorter frames in modelling sharp speech onsets.

The SB LPC vocoder, including its fixed and variable frame rate structures, outperformed the I-MBE standard during the tests due to a large difference in cleanness and smoothness between the two codecs as perceived by a considerable majority of the test subjects.

5. CONCLUSION

The proposed variable frame length vocoder represents a new approach to very low bit rate speech coding. Its main features are the source-dependent segmentation method at the encoder and the adaptive interpolation scheme at the decoder suited to the mechanism of the variable frame length vocoding. The vocoder is capable of preserving high quality speech whilst potentially reducing the fixed frame rate by 22% on average.

6. REFERENCES

[1] A. Kondoz, “Digital speech - Coding for Low Bit Rate Communication Systems”, Wiley, 1994

[2] DVSI (Digital Voice Systems Inc), “Inmarsat-M Voice Coding System Desrciption”, Draft 1.3, 1991

[3] E. Paksoy, K. Srinivasan, A. Gersho, “Variable Rate Speech Coding with Phonetic Segmentation” Proc. of IEEE ICASSP, pp. 155-158, 1993

[4] S. Villette, M. Stefanovic, I. Atkinson, A Kondoz, “High Quality Split Band LPC Vocoder and its fixed point real-time implementation” Proc. of EUROSPEECH, pp. 1243-1246, 1997