MULTIBAND PROTOTYPE WAVEFORM ANALYSIS SYNTHESIS FOR VERY LOW BIT RATE SPEECH CODING

K. Yaghmaie and A. M. Kondoz
CCSR, University of Surrey, Guildford GU2 5XH, UK

ABSTRACT

Prototype waveform interpolation is one of the most efficient compression techniques for coding the speech signal at bit rates below 4 kb/s. Most of the PWI coders employ prototype waveforms of the linear predictive residual signal for coding purpose. In the latest PWI systems, decomposition methods are used to separate the voiced and unvoiced components of the prototype waveforms prior to coding. This has resulted in high quality speech at very low bit rates. This paper presents a novel combination of the Multiband voicing analysis and PWI coding system in which the Multiband analysis is exploited to identify the voiced and unvoiced spectral components of the prototype waveforms of the original speech signal. To produce a high quality synthetic speech, energy variation of the original signal is recovered by transmitting its energy envelope. This method resulted in a high quality and low complexity coder operating at 2.55 kb/s.

1. INTRODUCTION

Prototype waveform interpolation (PWI) technique has proved to be one of the most efficient methods in coding the speech signal at very low bit rates [1] [2]. To achieve high quality speech at such low bit rates, the new PWI coders separate the voiced and unvoiced components of the linear predictive (LP) residual domain prototype waveforms prior to coding so that each component can be coded regarding its specific characteristics. In this group of PWI coders, decomposition methods based on the use of two dimensional Fourier transform is employed. The separated voiced and unvoiced components are then quantised independently. However, use of different bit patterns for voiced and unvoiced components, enforces use of a fairly large frame size which limits the performance of the coding system especially for high pitch speakers. Besides, the linear interpolation model used in such coders does not always preserve the energy variation of the original signal particularly for a long frame length and short pitch period.

The PWI method can be applied to the original speech signal too. In fact, in this case, very efficient quantisation methods may be applied regarding the high correlation of the adjacent prototype waveforms. Similar to the PWI in the residual domain, it is suitable to separate the voiced and unvoiced components of the speech signal before quantisation. While this can be achieved following a similar method, an alternative method is to classify the harmonics of the extracted prototypes as voiced and unvoiced.

This paper presents a method in which the Multiband analysis [3] is used to determine a frequency marker which separates the voiced and unvoiced spectral components of the speech domain prototypes. The scaled magnitude spectra of the prototype waveforms are encoded using a number of codebooks. The phase spectrum is simply modelled at the decoder side. The voiced components are synthesised as in conventional PWI coders. The unvoiced components, however, are reproduced by simply randomising the phase of the harmonics identified as unvoiced. This results in a significant saving in number of bits since only 4 bits are required to encode the location of the frequency marker. To remove the possible degradation in speech quality caused by inaccuracy of the Multiband model in the transition regions, additional information about voicing decision can be transmitted. To maintain the energy variation of the original speech, its energy envelope [4] is extracted and encoded for each frame. At the decoder, the recovered energy envelope is imposed on the output of the PWI coder. Also, to produce high quality speech, the frame size is limited to 20 ms.

The paper proceeds with a description of the basic coding algorithm in section 2. Quantisation of the prototype waveform and energy envelope is discussed in section 3 followed by a quality enhancement technique and simulations in sections 4 and 5. The results are discussed in section 6. The paper concludes with suggestions for further quality improvement and bit rate reduction as well as alternative approaches to implement the proposed idea.

2. CODING ALGORITHM

The algorithm can be described as a generalised PWI coding scheme which operates on the original speech signal. In the first stage, a Multiband voicing analysis is performed on each 20 ms speech frame to identify the voiced and unvoiced frequency bands (harmonics). The location of the marker which separates the voiced and unvoiced bands is our voicing measure, transmitted to the decoder as a frequency marker. In the case where there are some unvoiced bands between the voiced bands, the highest voiced frequency band is considered as the frequency marker. At the same time, for each speech frame a prototype waveform is extracted under minimum boundary energy criterion. Besides, the energy of the intermediate pitch cycles between

the current and next prototype waveforms are determined. These energy values are then normalised by the energy of the current prototype to produce the energy envelope of the current frame. The spectral magnitude of the prototypes and the energy envelopes are then quantised. No phase information is transmitted for the prototypes. Instead, at the decoder, the phases are set to a default value. This is reasonable since we always have an integer number of cycles between each two adjacent prototype waveforms. Furthermore, the auditory system of human is fairly insensitive to the phase characteristics of the speech signal. To produce the spectral amplitude of the missing intermediate cycles, the spectral amplitudes of the prototype waveforms are interpolated as in a conventional PWI coder. The time domain cycles are then produced by applying an inverse Fourier transform. In this process, the phase spectrum of the intermediate cycles are also set equal to those of the prototype waveforms i.e the default value. However, before applying the inverse transform, a phase jitter is introduced for the harmonics specified as unvoiced by the frequency marker. Figure 1 shows the block diagram of the encoder and decoder of the Multiband PWI system.

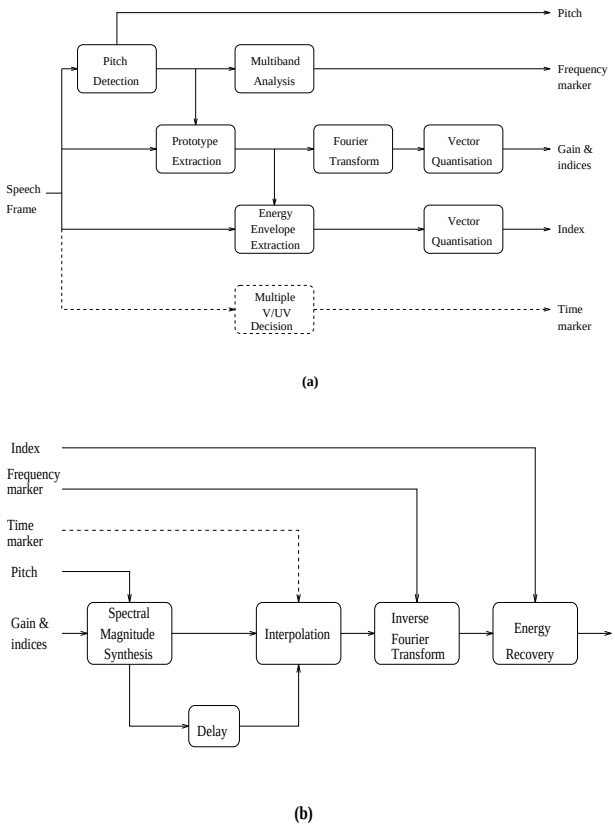

Figure 1. The block diagram of the (a) encoder and (b) decoder

3. QUANTISATION

The prototype waveforms spectra are normalised to have unity energy before quantisation. The resultant spectra, is then scaled to have equal number of harmonics before quantisation. This size is determined by the number of the harmonics of the prototype with the longest pitch period. Both linear or non-linear scaling can be applied [5] ,the former being simpler to implement while the latter being more efficient in term of quality. The scaled up spectrum of each prototype is then split into a number of overlapping sections each of which is vector quantised. Figure 2 shows the pattern for three overlapping regions as used in this work. The maximum number of harmonics is considered to be 74 corresponding to pitch period equal to 147 samples at 8 kHz sampling rate. A simple method to improve the matching when the linear scaling is used, is to consider only the scaled harmonics in the search process. This is because only such harmonics are selected in the down-scaling process at the decoder. Experiments showed that this simple measure leads to 2.1 dB increase in the average SNR between the quantised and unquantised spectra. Figure 3(a) and 3(b) show the quantisation and inverse quantisation processes at the encoder and decoder.

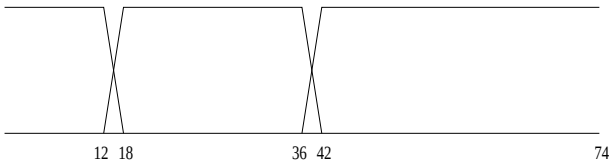

Figure 2. The frequency regions for quantising the extended prototype spectrum

The energy envelope of the frames are quantised following a similar approach. In this case, each energy envelope is scaled up to a vector with size 8, corresponding to the maximum number of cycles that can exist in a speech frame.

4. FURTHER QUALITY ENHANCEMENT

The above algorithm produces high quality speech at bit rates as low as 2.4 kb/s. However, the performance of the the coding system is affected by the inherent inaccuracy of the Multiband voiced/unvoiced analysis which especially appears in the transient regions of speech. This is mainly because such regions (especially the transients from unvoiced to voiced), have distinct portion of voiced and unvoiced sections in time domain which can not be modelled well by the Multiband model. In the experiments, it was found that adding a time domain marker to the described model can enhance the performance. Such time marker informs the decoder of the number of the purely unvoiced cycles in each transient frame so that the interpolation can be performed more accurately.

5. SIMULATIONS

The above method was employed to design a 2.55 kb/s MBPWI coder. Table 1 shows the bit allocation for the different parameters in this design. The expanded spectra

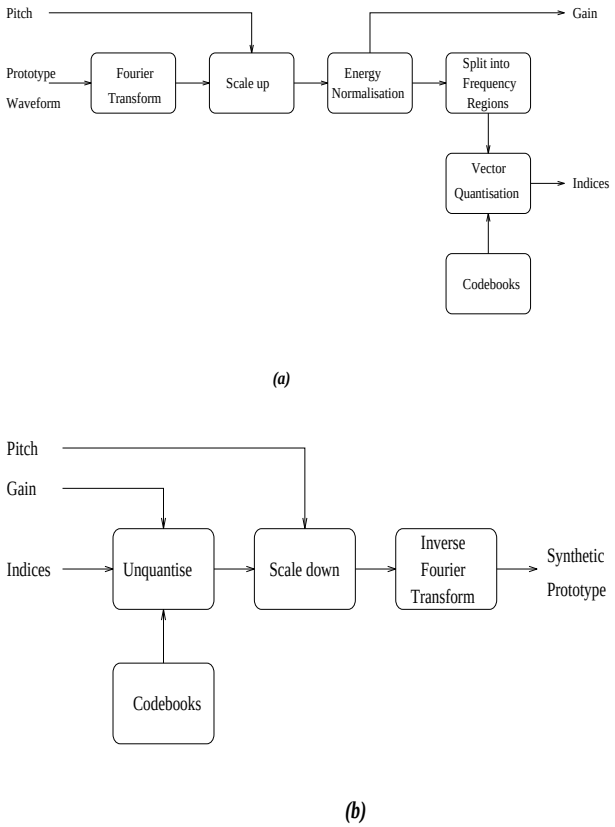

Figure 3. The block diagram of the (a) quantisation and (b) inverse quantisation processes

of the prototype waveforms were quantised using three 9 bit codebooks. The normalising factor (gain) for each prototype was quantised using a 5 bit scalar quantiser and a 5 bit codebook was employed to encode the scaled energy envelopes. The frequency and time markers are encoded using 4 and 3 bits respectively. To prevent discontinuities during the interpolation process, the gain values extracted from the energy envelope were interpolated before being imposed on the output signal of the PWI coder.

In the above implementation, all parameters were quantised in a non-differential scheme. This makes the coding system more robust to channel errors. Also to achieve high quality synthetic speech, the frame length was limited to 20 ms. Therefore, for a lower bit rate application an increase in the frame length can be considered.

6. RESULTS

To evaluate the performance of the above algorithm, computer simulations were conducted. Figure 4(a) shows a frame of a mixed type speech signal in which the prototype waveform is indicated by a window. Figure 4(b) illustrates the magnitude spectrum of this prototype together with the frequency marker which separates the voiced and unvoiced harmonics. The quantised spectrum is shown in figure 4(c). It can be observed that the real and synthetic spectra matches well especially in the voiced region. In

Parameters	Bits
Pitch period	7
Spectral codebooks	3*9
Spectral gain	5
Energy envelope	5
Frequency marker	4
Time marker	3
Total bits	51

Table 1. Bit allocation for the 2.55 MBPWI coder

an objective test, The average SNR between the real and quantised spectra found to be more than 21 dB. It is expected that by optimal selection of the frequency regions and size of codebooks, an even more accurate match can be achieved. The above SNR is slightly (about 10%) higher for the high pitch (female) speakers. This could be expected since for such speakers, less number of harmonics is involved in the search process, increasing the possibility of finding an accurate match. However, since for these speakers higher number of intermediate cycles is involved in the synthetic output speech, a fairly uniform quality can be expected for all speakers regardless of the pitch period.

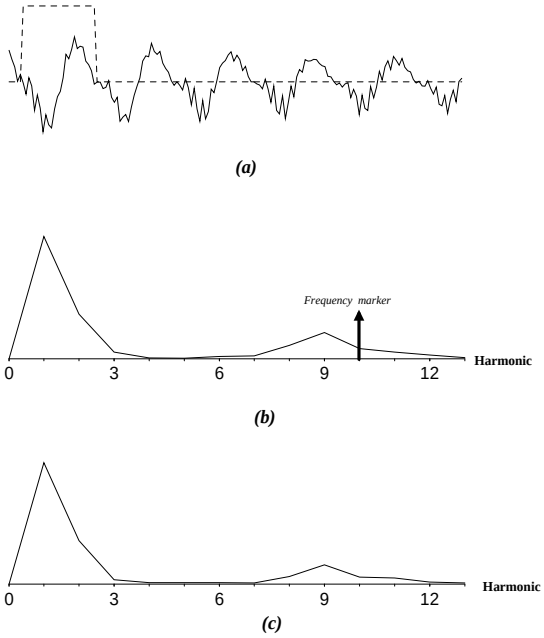

Figure 4. (a) A frame of mixed type speech signal (b) the spectra of the prototype signal with the pitch marker and (C) the spectra of the quantised prototype

In an informal subjective test, the coded speech of the 2.55 kb/s of the MBPWI outperformed those of IMBE [6] and FS1016 [7] coders. These tests also confirmed the consistent quality for different speakers discussed above. A slight degradation could be detected when the transient regions were detected wrongly. In such cases, the quality at

these regions is reduced to that of the IMBE coder. Further work is under investigation to develop a more accurate time marker detector algorithm.

7. CONCLUDING REMARKS

In this paper a low bit rate coding system based on prototype interpolation coding and Multiband voiced/unvoiced analysis were presented. To achieve this, the voiced and unvoiced spectral components are identified by a frequency marker determined in a Multiband voicing analysis. The unvoiced components are then simply produced by randomising their phase in the synthesis process. This simple method releases a significant number of bits required to encode the unvoiced components, enabling use of 20 ms speech frame. This leads to a higher performance especially for medium and high pitch speakers. To further improve the quality, the energy variation of the speech signal is recovered by encoding the energy envelope of the speech signal on pitch cycle basis.

In the presented method, the magnitude spectra of the prototype waveforms were coded using a number of overlapping codebooks. The frequency regions were not optimised, neither the number of bits per codebook. Therefore, it is expected that by such optimisation an improvement in performance and/or a reduction in bit rate can be achieved.

As an alternative method, the speech waveform can be split into purely voiced and unvoiced components using the same frequency marker concept (by simply filtering out the other component). The PWI can then be employed to encode the purely voiced components. In such an approach, a more efficient coding algorithm can be expected for encoding the spectra of the voiced prototypes. The saving on the bits may then be employed to encode the unvoiced components independently.

To maintain the simplicity of the algorithm, linear interpolation was used to extend the spectra of the prototype waveforms. However, if the additional computational work is acceptable, non-linear interpolation can also be employed to obtain a higher performance without any increase in bit rate.

For lower bit rate applications, a fixed frequency marker can be employed. Moreover, the time marker can be ignored and fewer bits can be employed for coding the energy envelope. These, together with a slight increase in the frame size and redesign of the involved codebooks can reduce the bit rate to below 2 kb/s. The preliminary experiments showed that at this rate, the coded speech is still highly intelligible and only suffers from slight buzziness. It was especially found that at these rates, employing a different strategy to encode the purely unvoiced frames results in a significant improvement in the performance.

Throughout this work, non-differential quantisation was used in all stages to make the whole coding system more robust to channel errors. However, for a low noise channel the differential coding may also be employed. This is expected to further reduce the bit rate because of the high similarity of the spectra of the adjacent prototype waveforms.

REFERENCES

[1] W. Bastiaan Kleijn. "Encoding Speech Using Prototype Waveforms". *IEEE transactions on speech and audio processing*, 1(4):386–399, October 1993.

[2] W. B. Kleijn and J. Haagen. "A speech coder based on decomposition of characteristic waveforms". *Proceedings of IEEE international conference on Acoustics, speech and signal processing*, pages 508–511, 1995.

[3] D. W. Griffin and J. S. Lim. "Multiband Excitation Vocoder". *IEEE Trans. on Acoustics, Speech and Signal Processing*, ASSP-36(8):1223–1235, 1988.

[4] A. M. Kondoz I. A. Atkinson and B. G Evans. "Time Envelope Vocoder, a New LP Based Coding Strategy for Use at Bit Rates of 2.4 kb/s and Below". *IEEE Journal on selected areas in communications*, 13(2):449–457, 1995.

[5] Paul C. Meuse. "A 2400 bps Multiband excitation vocoder". *Proceedings of IEEE international conference on Acoustics, speech and signal processing*, pages 9–12, 1990.

[6] Digital Voice System Inc. Inmarsat M. *"System Definition Manual(SDM) Module 1 Inmarsat Voice Codec"*, August 1991.

[7] V.C. Welch J.P.Campbell and T.E. Tremin. "The DOD 4.8 kb/s Standard (Proposed Federal Standard 1016)". *B. S. Atal, V. Cuperman and A. Gresho, editors, "Advances in Speech Coding"*, pages 121–123, 1991.