PITCH SYNCHRONOUS MULTI-BAND (PSMB) SPEECH CODING

Haiyun Yang Soo-Ngee Koh Pratab Sivaprakasapillai

School of Electrical and Electronic Engineering
Nanyang Technological University, Nanyang Avenue, Singapore 2263
ehyyang@ntuvax.ntu.ac.sg

ABSTRACT

A novel speech coding algorithm, named pitch synchronous multi-band (PSMB), is proposed. It uses the multiband excitation (MBE) model to generate a representative pitch-cycle waveform (PCW) for each frame. The representative PCW of a frame is encoded by two out of three codebooks depending upon whether the frame is related or unrelated to the previous frame. The new speech coder introduces a pitch-period-based coding feature. The PSMB coder operating at 4 kbps outperforms the Inmarsat 4.15 kbps IMBE coder by a clear margin. It is also found to be slightly better than the FS1016 4.8 kbps code excited linear predictive (CELP) coder in terms of perceptual quality. Fast search algorithms for the three codebooks used in PSMB are also developed.

1. INTRODUCTION

A common feature of the code-excited linear predictive (CELP) coders [1] is its frame-based procedure. However, the basic unit of a voiced speech signal waveform is the pitch period in duration. It therefore seems more efficient to consider the pitch-period-based approach instead of the conventional frame-based approach in designing a speech coder. The new speech coder includes the pitch-period-based feature. It concentrates on the processing of the PCWs. To achieve a fixed bit rate coding, most of the parameters are estimated using a representative PCW generated for each analysis frame. The multi-band excitation (MBE) model [2] is used to generate the PCW. After efficiently encoding the PCW, speech signals are synthesized using the method described in [3].

For adopting the pitch-period-based feature, the PSMB coder is similar to the prototype waveform interpolation (PWI) coder [4]. However, there is a major difference. The PSMB coder encodes both voiced and unvoiced signals whereas the PWI coder encodes voiced signals only. Although PCW is only meaningful for voiced speech signals, the method of generating PCWs and also the scheme of encoding PCWs (to be discussed shortly) make it suitable for the new speech coder to process unvoiced speech signals as well. The PSMB coder is also related to the MBE model because PCWs are obtained by using multi-band analysis. However, some of the weaknesses of the MBE model have been overcome [5] in the proposed coder by encoding the PCW with a closed-loop analysis-by-synthesis (ABS) structure.

2. PSMB SPEECH CODER

The basic structure of the PSMB coder is presented in Fig.1. A refined pitch period which is obtained as described in [3], is used to divide the spectrum into several bands, each of which encompasses one harmonic. The voiced/unvoiced (v/uv) decision is made for each group of three bands. A single PCW can be generated as :

$$w(i) = \sum_{j=1}^{L} M_j \cos\left(2\pi \frac{ij}{P} + \theta_j\right) \quad i = 0, 1, ..., P-1 \quad (1)$$

where M_j is the magnitude and θ_j is the phase of j-th band, P is the integer part of the refined pitch period and L is the number of bands. For simplicity, the integer part of the refined pitch period will be referred to as the pitch period.

The PCW generated by Eq.(1) is actually the representative PCW of a frame. Once the synthesized PCW, denoted by $\hat{w}(i)$, is obtained, the magnitudes and phases of all bands are reconstructed as follows :

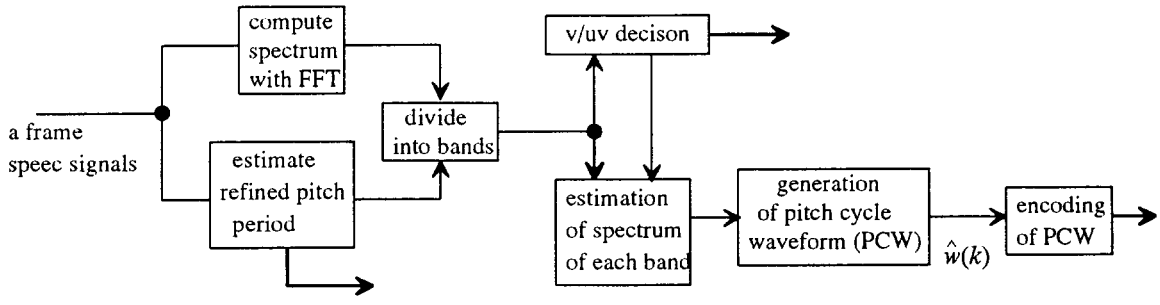

Figure 1 Structure of Pitch-Synthronous Multi-Band Coder

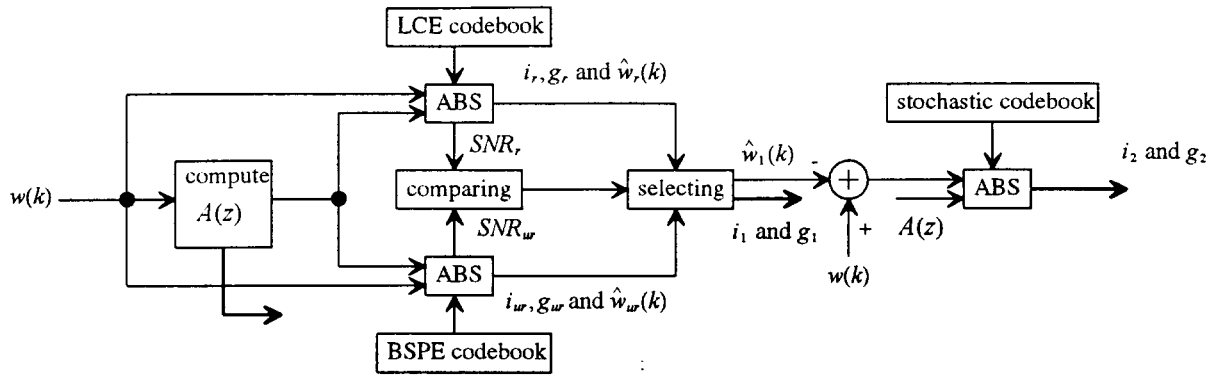

Figure 2 Basic Structure of Encoding PCWs

$$\hat{M}_j = \frac{2}{P}\left\{\left[\sum_{i=0}^{P-1}\hat{w}(j)\cos\left(2\pi\frac{ij}{P}\right)\right]^2 + \left[\sum_{i=0}^{P-1}\hat{w}(j)\sin\left(2\pi\frac{ij}{P}\right)\right]^2\right\} \quad (2)$$

$$\hat{\theta}_j = -\arctan\frac{\sum_{i=0}^{P-1}\hat{w}(j)\sin\left(2\pi\frac{ij}{P}\right)}{\sum_{i=0}^{P-1}\hat{w}(j)\cos\left(2\pi\frac{ij}{P}\right)} \quad (3)$$

where j = 1, 2, ..., L-1. Next, the voiced bands are synthesized using a method similar to that described in [3]. The true phase of each voiced band is used in the synthesis procedure instead of a reconstructed phase obtained through phase smoothing [3]. For unvoiced bands, the synthesis method described in [3] is used.

2.1. Encoding of PCW

For voiced speech signals, the PCWs of two neighbouring frames are in general highly correlated. However, during the onset of speech signals or the transition periods, the PCWs change rapidly in their lengths and shapes. To treat these two cases differently, a related/unrelated (r/ur) decision is used for each frame. The method of making the r/ur decision is depicted in Fig.2, which presents the encoding of PCWs. To encode a PCW, its LPC coefficients are computed first. Two ABS structures are used in parallel, one corresponding to the assumption of the current frame being a related frame and the other to it being an unrelated frame. The corresponding codebooks, to be described later, are the length-converted-excitation (LCE) codebook and the bandlimited-single-pulse-excitation (BSPE) codebook. Processing through each of the two parallel ABS structures, produces a signal to noise ratio (SNR), an optimal index (i), a gain value (g), and a synthesized PCW ($\hat{w}(k)$). The "comparing" block makes the r/ur decision, depending on which of the two SNRs (SNR_r and SNR_{ur}) has the higher value. The "selecting" block chooses the first-step synthesized PCW, $\hat{w}_1(k)$, from $\hat{w}_r(k)$ and $\hat{w}_{ur}(k)$, according to the r/ur decision. For both related and unrelated frames, the error signal between $w(k)$ and $\hat{w}_1(k)$ is a noise-like signal. Another ABS structure using a stochastic codebook is used to encode the error signal. Through a closed-loop procedure, the optimal codevector, which produces the minimum mean-square error, is selected.

2.2. Circular-Analysis-Filtering and Circular-Synthesis-Filtering with LPC

At some of the stages of the PSMB coder, it is necessary to compute the residual signal of a PCW. Let $\{a(i) : i = 0, 1, ..., N$, where $a(0) = 1.0\}$ be the LPC coefficients of the PCW $w(k)$, where N is the order of the LPC coefficients. The residual signal can be computed by

$$e(i) = \sum_{k=0}^{P-1} a(k)w\left((i-k)\right)_P \quad (4)$$

where $((\ n\ ))_P$ denotes the (n[mod P]) operation. For convenience, the residual signal $e(i)$ is named as the excitation PCW. The analysis-filtering expressed by Eq.(4) is named as circular-analysis-filtering since circular convolution is used here.

Conversely, it is also necessary to obtain the PCW from its excitation PCW. This is achieved as,

$$w(i) = \sum_{k=0}^{P-1} d(k)e\left((i-k)\right)_P \quad (5)$$

where $d(k)$ is the P-order inverse DFT (IDFT) of the expression $\frac{A(k)}{|A(k)|^2}$, k = 0, 1, ..., P-1, and $A(k)$ is the P-order DFT of the LPC coefficients [5]. Because Eq.(5) is also a circular-convolution, the filter $d(k)$, where k = 0, 1, ..., P-1 is named the circular-synthesis-filter

The benefit of implementing the circular synthesis-filtering process in the time, instead of frequency [5] domain is that it can provide a fast algorithm in the search procedure.

2.3. Related Frame Processing and LCE Codebook

For the processing of related frames, an excitation PCW is first obtained by circular-analysis-filtering of the PCW of the previous frame. The length of the excitation PCW is converted to that of the current PCW while maintaining its shape as far as possible [5]. Next, the length converted excitation PCW is circularly shifted to match the the excitation PCW obtained from the current

PCW. Finally, a fixed gain coefficient is used to modify the magnitudes.

The LCE codebook is used to achieve the above processing procedure in the ABS structure. This codebook consists of different rotations of the LCE vector which is the length-converted excitation PCW of the previous frame. The length of all codevectors is the pitch period of the current frame. To design a given size of LCE codebook, 64 (say), we have to construct, as uniformly as possible, 64 individually rotated vectors of the LCE vector. If the pitch period is greater than the duration of 64 samples, 64 circular integral-shifts of the LCE vector can be chosen from all possible circular integral-shifts to make up the LCE codebook. However, when the pitch period is less than the duration of 64 samples, some fractionally rotated versions of the LCE vector have to be included in the LCE codebook because the number of all possible integer rotations is still less than 64, the necessary number of codevectors. Furthermore, to make the number of codevectors exactly equal to 64 in this case, two values of resolution have to be generally used. The rotations corresponding to some of the codevectors will have one value of the resolution while those corresponding to the remaining codevectors will have the other value. This information has to be pre-stored.

2.4. Unrelated Frame Processing and BSPE Codebook

It is known that speech signals can be synthesized by passing an excitation signal through a linear system which models the vocal tract. For voiced speech signals, the excitation source is usually represented by a periodic pulse train, where the spacing between the consecutive pulses is the pitch period. With this assumption, a PCW can be reconstructed by passing through the linear system, a bandlimited single pulse excitation signal which is suitably located within a pitch period. Since the pulse may appear at any position during one pitch period, the BSPE codebook, whose codevectors are obtained by rotating a pre-stored bandlimited single pulse vector, is used. The rotation may be of integral or fractional resolution. As in the case of the LCE codebook, the BSPE codebook generally has more than one fractional resolution for some pitch periods. The corresponding information also needs to be pre-stored [5].

2.5. Synthesis of PCW

The reconstructed PCW of the current frame consists of two parts. One is a synthesized component either from the LCE codebook or BSPE codebook, which is decided by the r/ur information. The other part is the synthesized component from the stochastic codebook.

3. 4.0 KBPS PSMB SPEECH CODER

To investigate the PSMB coder further, a 4.0 kbps PSMB coder is developed. It is outlined first in this section. The experimental results obtained from the simulations of the coder are then given. The computational complexity of the coder is also discussed.

3.1. Quantization of Parameters

The bit allocation for the coder parameters discussed in Section 2, is given in Tab.1. The pitch period and the v/uv decisions are directly quantized [3]. The 10-order LPC coefficients are converted to line spectrum pairs (LSP) [6], and differentially quantized. For related frames, the 6-bit LCE codebook is accurate enough, and its corresponding stochastic codebook size is 512. The two gains are linearly and logarithmically quantized. For unrelated frames, the sizes of the BSPE codebook and stochastic codebooks are 128 and 256 respectively. Both gain coefficients are then logarithmically quantized. All parameters are updated every 20 ms.

The speech signals in the TIMIT corpus CD-ROM were used to train the quantizers. The TIMIT speech signal was low-pass filtered by a 240-order FIR filter and down-sampled to 8 kHz. A total of 434 speech utterances, spoken by 105 females and 112 males, made up the training data set.

Table 1 Bit Allocation of 4.0 kbps PSMB Coder

Parameters	Bit	Bit Rate (kbps)
pitch period	8	0.4
v/uv	12	0.6
LPC	33	1.65
r/ur	1	0.05
LCE or BLSP and stochastic codebook	15	0.75
g_1	5	0.25
g_2	6	0.3
Total	80	4

3.2. Experimental Results

The processing results, using the 4.0 kbps PSMB coder, for a voiced speech frame is presented in Fig.3. As shown in Figs.3(a) and (b), the synthesized speech waveform is very similar to that of the original signal, and the synthesized signal is synchronous with the original one because of the inclusion of phase information through the novel idea of PCW coding. For unvoiced speech signals, because randomly generated phases are used while synthesizing unvoiced bands, the synthesized unvoiced speech waveform does not look like the original speech waveform. However their spectral envelopes are still very similar to one another [5].

The testing data set in our experiments consisted of 22 speech utterances from the TIMIT database, not previously selected for the training data set. They were spoken by 5 males and 7 females. The average signal-to-noise ratio (SNR) value of the representative PCWs of the testing data set was 9.07 dB for the 4.0 kbps PSMB coder, while the average SNR for the test data set was 7.55 dB for the FS1016 standard CELP coder operating at 4.8 kbps [7]. Although, as a performance measure, the SNR of PCWs is not the same as the SNR of decoded speech waveforms, the 9.07 average SNR of PCWs implies that the coding of PCWs presented in this paper is quite efficient.

To assess the decoded speech quality of the new coder, a subjective exercise was performed by eleven listeners. The subjective

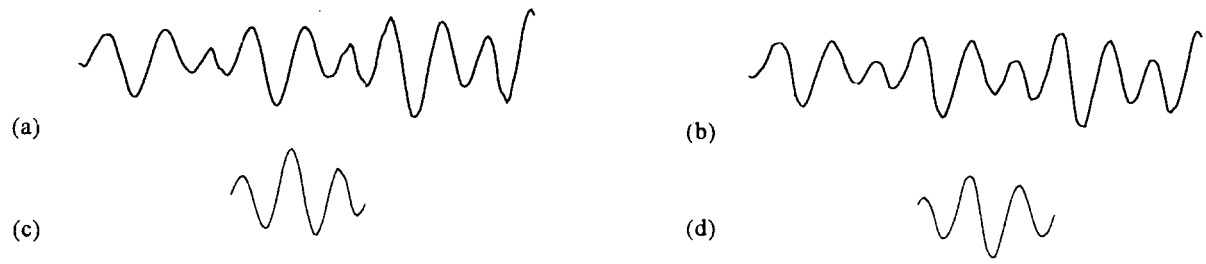

Figure 3 (a) Voiced Speech Frame (20 ms) (b) Synthesized Speech Frame
(c) Representative PCW (Original) (d) Representative PCW (Synthesized)

evaluation included two groups of pair comparison tests. The first group was designed to compare the PSMB speech coder at the bit rate of 4.0 kbps with the 4.15 kbps Inmarsat standard IMBE [3]. The second group was designed to compare the same PSMB speech coder with the FS1016 CELP. The subjective evaluation results are presented in Tab.2.

Table 2 Subjective Evaluation through Pair Comparisons

PSMB vs IMBE

Preference	for PSMB	Uncertain	for IMBE
No. of cases	141	63	38

PSMB vs CELP

Preference	for PSMB	Uncertain	for CELP
No. of cases	92	85	65

It can be concluded from the subjective evaluation that the 4.0 kbps PSMB coder performs significantly better than the 4.15 kbps Inmarsat IMBE coder. It was also observed that the PSMB coder does not suffer from any hoarseness that is sometimes present with the IMBE coder [5]. In addition, it is found that the PSMB coder is slightly better than the 4.8 kbps FS1016 CELP coder. When comparing the CELP coder with the PSMB coder through informal listening tests, it was observed that the two coders have slightly different audible distortions. The speech signals decoded by PSMB sounded slightly smoother that those synthesized by CELP. Also, the decoded speech from the CELP coder was found to be slightly unnatural for some of the utterances. On the other hand, some utterances from the CELP coder were slightly cleaner perceptually than those from the PSMB coder. This seems to be the main reason for some of the utterances from CELP being judged as better than those from PSMB.

3.3. Computational Complexity

In speech coders with the closed-loop ABS structure, the high computational load arises mainly from the selection of the optimal vector from codebooks. There are three codebooks in PSMB. A fast search algorithm is developed for each codebook [5], which reduces the searching computational complexity of the 4.0 kbps PSMB coder to 4.54 mips [5]. Considering the other operations included in the coder, its total computational complexity is estimated at 12.2 mips. This is achievable by currently available DSP chips.

4. CONCLUSION

Our experiments have shown that the proposed 4 kbps PSMB coder is significantly better than the IMBE coder in terms of the perceptual quality of the decoded speech. It has also been found to be slightly better than the 4.8 kbps CELP coder in subjective performance. Also, for the coding of PCWs, the proposed scheme using two codebooks based on the r/ur decision has been proved to be very efficient. The fast search algorithms developed for the three codebooks reduce the computational complexity of the overall coder to a level comparable to that of the FS1016 CELP coder. It is therefore possible to implement the proposed coder for real-time applications, using currently available DSP chips.

Reference

[1] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP) : high-quality speech at very low bit rates", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp.937-940, 1985.

[2] D. W. Griffin and J. S. Lim, "Multiband excitation", IEEE Trans. Acoust., Speech, Signal Process. ASSP-36, pp.1223-1235, 1988.

[3] Digital Voice Systems Inc., "Inmarsat-M voice codec", ver.3.0, August 1991.

[4] W. B. Keijin, "Encoding speech using prototype waveforms", IEEE Trans. Speech and Audio Process. vol.1, No.4, pp.386-399, October 1993.

[5] H. Yang, S. N. Koh and P. Sivaprakasapillai, "Pitch -synchronous multi-band (PSMB) coding of speech signals", to be published.

[6] F. Soong and B. H. Huang, "Line spectrum pair and speech data compression", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, paper 1.10, 1984.

[7] R. Fenichel, "Proposed federal standard 1016 (pre-third draft)", National Communications System, Office of Technological Standards, Washington DC 20305-2010, 8 July 1990.