A 1.7Kbps Waveform Interpolation Speech Coder Using Decomposition of Pitch Cycle Waveform

Ge Gao and P.C. Ching
Department of Electronic Engineering,
The Chinese University of Hong Kong, Shatin, Hong Kong
Tel:(852)26098275 Fax:(852)26035558
email: {ggao, pcching}@ee.cuhk.edu.hk

ABSTRACT

In this paper, we propose a low bit rate waveform interpolation speech coder where the novelty lies with an effective decomposition method of pitch cycle waveform (PCW). PCWs exhibit very different perceptual characteristics in different frequency bands. For frequency components below 1kHz, they are quantized using Variable Dimensional Vector Quantization (VDVQ) scheme. Hereby retaining the fine harmonic structure of the speech signal. For the upper frequency band ranging from 1 to 4kHz, the formant structure is perceptually more dominant. It is therefore desirable to capture the formant peaks in order to maintain high speech quality. In this case, we employ a sparse frequency representation (SFR) method. A 1.7kbps speech coder has been developed by this technique and the quality of the output speech is perceptually good.

INTRODUCTION

It has been demonstrated that waveform interpolation speech coder can produce high quality synthetic speech below 4kbps [1]. In this approach, voiced sound is generated by interpolating successive Pitch Cycle Waveforms (PCW). Due to the quasi-stationary characteristics of voiced speech, the pitch cycle waveform (PCW) usually evolves very slowly. Thus it is possible to update the PCW at every 20 to 30 ms interval. An important step in waveform interpolation technique is the decomposition of PCW [1]. Unvoiced and voiced speech are normally treated the same and an arbitrary period will be assigned to any unvoiced segment. Low pass and high pass filters are applied to obtain the Slow Evolving Waveform (SEW) and Rapid Evolving Waveform (REW) respectively. SEW represents the periodically changing voiced portion whilst REW gives the noisy part. In the decoder, SEWs are interpolated and added to REW to recover the speech signal. However, it might be more efficient to treat the voiced and unvoiced speech separately since they have totally different acoustic characteristics. It is well-known that unvoiced speech can be generated by passing white Gaussian noise through a LPC synthesis filter. Whereas a high correlation between successive PCWs is observed among voiced segments. This observation leads to differential coding or interpolation coding. However, differential coding [2] is sensitive to channel errors due to its error propagation properties. It is advantageous to code the PCW directly. The PCW can be generated by exciting the LPC synthesis filter with LPC residue. Human ears have different perceptual response to different frequency bands. The harmonic structure is important for lower frequency band while formant information is indispensable at higher frequency. This difference motivates us to decompose the PCW into different frequency bands and code them separately. For frequency components below 1kHz, we apply DCT to encode the signal. The DCT coefficients are then quantized with the technique Variable Dimensional Vector Quantization (VDVQ). For frequency band in the range of from 1kHz to 4kHz, we retain only the formant information.

A new method of generating PCW is investigated in this paper. The decomposition of PCW into SEW and REW incurred a large coding delay and high computation requirements. The alignment of the oversampled PCW also has a high computational load. PCW produced by this technique has the advantage of low coding delay and low computation. The unquantized version of this coder gives very high quality speech. We have then developed a 1.7Kbps speech coder in which the output quality is still very much comparable to codecs operate at 4kbps.

OVERVIEW OF THE CODER

Incoming speech is segmented into frames and each frame is further divided into two subframes. Each subframe will then be labeled as voiced or unvoiced independently. During stationary voiced seg-
ment, speech can be recovered from the interpolation of speech PCW. Onset and offset regions are recovered through a weighted sum of unvoiced and voiced portion; and unvoiced speech is recovered by exciting the LPC synthesis filter with white Gaussian noise having an appropriate energy envelope. This sub-frame division has the merit of retaining a lower update rate during voiced and unvoiced speech segments while still capable to capture fast changes around onset and offset. Smooth transition during onset and offset is also guaranteed. Onset regions often come up with a sharp increase in energy since offset regions have a tapered energy envelope. To overcome this we use different weighting functions for these two situations.

Pitch estimation and voicing state decision is crucial to speech quality. Any error will essentially give rise to serious perceptual degradation. We have therefore developed a very reliable, robust pitch estimation algorithm with voicing state decision. This enables the proposed speech coder to work satisfactorily even under noisy environment. Details of this pitch estimation algorithm and voicing state decision criteria will be reported elsewhere.

![Figure 1: Spectrum of the LPC residue](image)

**GENERATING PCW**

There are several methods to generate the PCW. In [1], over sampled PCWs are first aligned and then fed through a low-pass filter. PCW is obtained by downsampling thereafter. In [2], PCW is extracted near the frame boundary where maximum correlation value can be found. We, however, use a different method for our codec. First of all, LPC coefficients are estimated from each incoming voiced segment. Current speech frame is then passed through LPC analysis filter to obtain the LPC residue. The residue signal is hammed windowed and padded with zeros to give 512 points. DFT is applied in which only amplitude information of the resulting spectrum is retained. All frequency components are assumed to have zero phase. A search method similar to the spectral envelope estimation vocoder(SEEVO) is applied to extract the peaks of the amplitude spectrum. In an ideal situation, the magnitude should be flat within the entire frequency range under consideration. Practically it is not flat. In the new American 2400bps standard [5], this spectral envelope is quantized with a 8-bit codebook. We apply this excitation to the LPC synthesis filter to get the PCW. It is not quantized directly. Figure 1 illustrates this procedure. The symbol $^*$ represents the peak found during the search process. The number of peaks should be the same as the total number of harmonics within the Nyquist frequency. Figure 2 is one period of the synthesized LPC excitation. Now assuming a voiced segment is periodic then the speech signal can be represented by:

$$s(n) = \sum_{k=0}^{N-1} A_k(n) \cos k\omega n + \sum_{k=1}^{N-1} B_k(n) \sin k\omega n. \quad (1)$$

While one period of the LPC residue can be written as

$$r(n) = \sum_{k=0}^{N-1} C_k(n) \cos k\omega n. \quad (2)$$

where $\omega$ is the fundamental frequency and $N$ is the total number of harmonics within Nyquist frequency. We can compute $A_k$ and $B_k$ from $C_k$:

\[
\begin{align*}
A_k(n) &= \frac{C_k(n) \sum_{i=0}^{M} a_i \cos k\omega i}{(\sum_{i=0}^{M} a_i \cos k\omega i)^2 + (\sum_{i=0}^{M} a_i \sin k\omega i)^2}, \\
B_k(n) &= \frac{-C_k(n) \sum_{i=0}^{M} a_i \sin k\omega i}{(\sum_{i=0}^{M} a_i \cos k\omega i)^2 + (\sum_{i=0}^{M} a_i \sin k\omega i)^2},
\end{align*}
\]

$k = 0, 1, \ldots, N - 1$

where $a_i$ is the LPC coefficients and $M$ is the prediction order.

![Figure 2: Synthesized LPC residue(all the phase set to zero)](image)
This PCW can be regarded as the ‘average’ PCW for the current frame since both the LPC synthesis filter and LPC residue are obtained from the entire frame instead of a single cycle of speech. This scheme avoids the problem of choosing an individual pitch period waveform which might not give the best representation particularly during the transition between two vowels. Alignment and filtering operation for oversampled PCWs as in [1] are not needed. Another advantage of the proposed method is that bandpass voicing can be used for higher bit rate application, which could improve the naturalness of speech. Compared with other coders that extract PCW directly from original speech frames, subjective listening test shows that synthesized speech obtained by this method produces better speech quality.

**ENCODING PCW**

Direct encoding of PCW requires a large number of bits. It has been reported [2] that Discrete Cosine Transform (DCT) can efficiently compress the PCW provided that the lower 30 percent coefficients are carefully quantized. While the remaining coefficients can be represented by a single value, it is not quite satisfactory in most cases since the high frequency formants play an important role in the intelligibility of speech. A single value to represent the energy of high frequency component will inevitably degrade speech quality noticeably. It is possible to transmit formant information to the receiver and in general only the position and amplitude are required. This is because human ears are not sensitive to the exact frequency of formant, thus we can quantize the position of the formant frequency coarsely. In [4], a sparse frequency representation of LPC residue is proposed where one cycle of the LPC residue is expressed as:

\[
\hat{u}(n) = \sum_{k=0}^{K} a_k w_k(n) \cos \omega_k n + \sum_{k=1}^{K} b_k w_k(n) \sin \omega_k n.
\]

(4)

Here, K is the number of basis functions selected for the sparse frequency representation (SFR) \(K \leq N\); \(w_k(n), k = 0, \ldots, K\), are the window functions; and \(a_k, b_k\) are the coefficients for the sparse representation. The frequency spacing is uniformly divided at frequency below 1kHz while logarithmically spaced at frequency above. In order to ensure that the whole frequency band will be covered, a window function (normally hamming window) is applied to these logarithmically distributed frequencies. This window function can ‘flatten’ the spectrum and they can be represented as follows:

\[
\omega_k = \begin{cases} 
    k \omega_0, & 0 \leq \omega \leq \omega_M \\
    \alpha \omega_{k-1}, & \omega_M \leq \omega_k \leq \omega_L
\end{cases}
\]

(5)

It can be noted that the window duration also decreases with an increase in \(\omega_k\). This scheme produces very high quality speech. However, it is not suitable for low bit rate speech coding because there are too many parameters to be encoded. It is, however, possible to apply this representation directly to the original speech signal to improve efficiency, yet there is no need to transmit the LPC coefficients. Formant information can be preserved. In this case, we select and encode eight frequency components between the range of 1 to 4kHz. The distribution is uneven and there are more representations in the low frequency range than in the high frequency. This ensures that no perceptual degradation is made due to quantization as human ears are less sensitive to the exact frequency of the formants in high frequency than in low frequency.

Formant information is estimated from the original LPC spectrum. In order to ensure accurate formant estimation, we use sample selective linear prediction (SSLP) [6] to estimate the LPC coefficients. Traditional autocorrelation method often results in the obscured formant peaks whilst pitch synchronous LPC analysis would have stability problems for very short pitch period. Since our coder depends heavily on the formant information, SSLP can give more accurate information about the formant peaks. A window function is applied to flatten the spectrum to spread power to the entire frequency range. Specifically, among the 8 predefined frequencies \(f(n), n = 1, 2, \ldots, 8\), if any one falls onto frequency \(i\) with amplitude \(A\), then we put \(f(i) = K \cdot A\), where \(K\) is coefficient determined by the pitch value. Longer pitch period usually requires larger \(K\). Since in a particular formant regions, they might contain more harmonics, thus more energy will be required. In the decoder, the PCW is synthesized by adding the lower frequency constituent to the higher one. It can be seen that phase information has been lost in the upper frequency range. We have actually assigned random phases to the formant peaks and this allows a saving of bits. Informal listening test shows that the speech quality has very little degradations at all.

For PCW below 1kHz, we apply DCT to it and use split vector Variable Dimensional Vector Quantization (VDVQ) to quantize the coefficients. For synthesis, PCW can be generated according to the following formula,

\[
s(n) = \sum_{i=0}^{K} a_i w_i(n) \cos \omega_i n + \sum_{k=1}^{K} b_i w_i(n) \sin \omega_i n,
\]

where \(K\) is now the number of harmonics below 1kHz plus 8, the number of sparse frequencies. Although the 8 frequency components are fixed, we cannot use them directly in the synthesizer. Because if they are not the exact harmonics of the fundamental frequency, discontinuities will occur on the boundaries of PCW. These discontinuities are perceptually very disturbing. We find that rounding these 8 fixed fre-
frequencies to the nearest harmonics gives a compromise solution. On one hand we do not need to transmit extra information while on the other hand maintain the continuity of the PCW. Figure 3 shows an example of the spectrum of the original wave compared with the reconstructed spectrum. We can see that formant information is preserved except that the frequency location has been shifted slightly, which, however, will not give rise to significant perceptual degradation.

Table 1 shows the bit allocation of this speech coder. There are totally four modes of speech frames to be classified according to their subframe characteristics. So two bits are required to represent them. During on-set regions, the parameter set to be transmitted to the receiver is the same as the voiced frame and the parameters for the unvoiced part is repeated from its previous frame; whereas during offset regions, the parameter set transmitted to the receiver is the same as the unvoiced frames and parameters for voiced part are repeated from the previous frame. This speech coder operates on a frame interval of 20ms, producing overall rate of 1.76kbps. It requires a look ahead of one and a half frame to ensure reliable pitch estimation and smooth transition between unvoiced and voiced frame.

The generation of PCW provides a suitable interface for the application of band pass voicing in the context of waveform interpolation. It can improve the naturalness of speech provided that more bits are available, and it does not involves the rigid form of REW and SEW. This is because unvoiced and voiced bands can occur at any frequency regions but not necessarily restricted as was described in [1].

CONCLUSION AND FUTURE WORK

A low bit rate speech coder based on waveform interpolation has been developed. An efficient and effective decomposition of PCW is suggested that ties closely to the inherent human acoustic characteristics. The proposed way of generating PCW also provides a new perspective to waveform interpolation speech modeling. We are now investigating the merits of this method on improving the performance of an interpolation based speech coder.

References


