HARMONIC CODING OF SPEECH AT 4.8 KB/S

Edward C. Bronson, Douglas A. Carlone, W. Bastiaan Kleijn, Kevin M. O'Dell, Joseph Picone, David L. Thomson

AT&T Bell Laboratories, Naperville-Wheaton Road, Naperville, Illinois 60566

ABSTRACT

This paper describes a new speech coding technique which yields improved speech quality over existing 2.4 kb/s LPC vocoders. The method is computationally efficient and operates at a data rate of 4.8 kb/s. Each speech frame is initially classified as voiced or unvoiced. Unvoiced frames are synthesized using a linear predictive coding filter with noise or multipulse excitation. Voiced frames are synthesized using a sum of sinusoids. The frequency of each sinusoid is defined by peaks in the frequency spectrum. A new interpolation technique provides a computationally efficient method of locating the spectral peaks. A real-time, fully quantized version has been implemented in hardware.

INTRODUCTION

In this paper, a 4.8 kb/s harmonic coding algorithm is presented which represents an improvement in speech quality over existing 2.4 kb/s LPC vocoder systems [1,2]. A major factor guiding algorithm design is that the algorithm be implementable in currently available hardware. Experience has shown that raising the bit rate of LPC vocoders by increasing the frame rate, the LPC filter order, and the quantization precision of the transmitted parameters does not lead to an audible improvement in speech quality. Thus, an alternative approach is desired. The Code Excited LPC method [3] provides very good speech quality, but, even in its efficient forms [4], is computationally too expensive. Most methods using sinusoidal reconstruction of the speech signal are oriented towards an 8 kb/s rate [5,6], and are also computationally very expensive.

The design of the 4.8 kb/s harmonic coder begins with an existing 2.4 kb/s vocoder [1] which consumes only a small part of the real-time available on the proposed hardware. The two main impairments of a typical LPC vocoder are the overall "buzzy" quality of the synthetic voiced speech and the lack of a robust voiced-unvoiced decision. The latter impairment is improved by utilizing a three-level voicing decision [7]. Here, we present a computationally inexpensive method to improve the quality of the voiced speech using harmonic reconstruction.

ANALYSIS

Fig. 1 diagrams the information flow between the major components of the analysis portion of the speech coder.

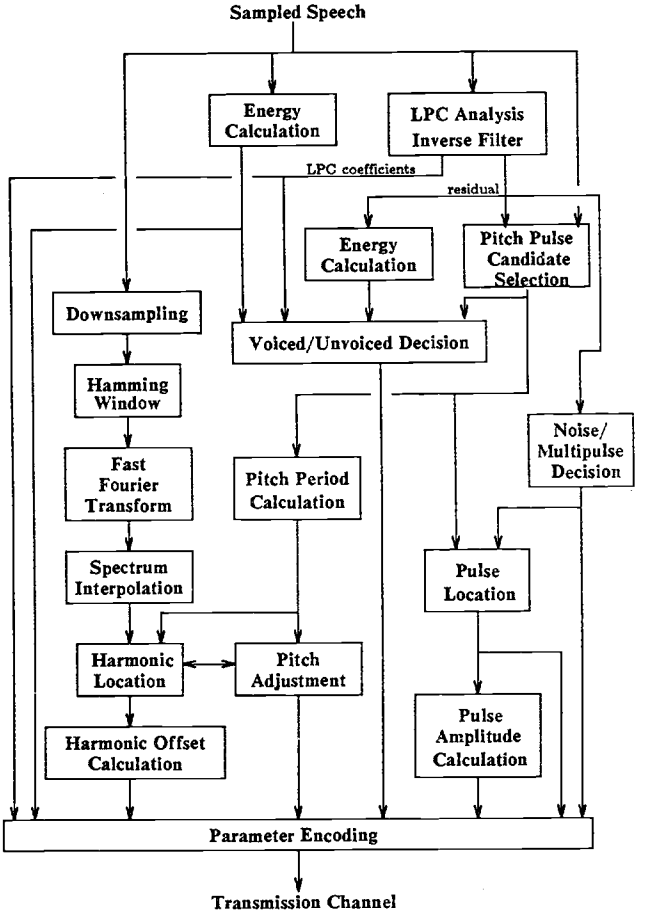

Fig. 1
Analyzer Block Diagram

The input analog speech is digitized at a sampling rate of 8 kHz and segmented into 20 ms frames. The Burg method [8] of LPC analysis is used to compute 16 LPC

CH2396-0/87/0000-2213 $1.00

coefficients since it is robust when using fixed-point arithmetic. Pitch detection, implemented using a time-domain discriminant analysis-based system [1], is performed prior to harmonic analysis.

The spectrum of a voiced frame consists of a series of peaks with frequencies that are close, but not exactly equal, to integer multiples of the speaker's fundamental frequency. Voiced frames are synthesized by summing sinusoids at selected frequencies in the neighborhood of multiples of the fundamental frequency. The analysis of a voiced frame uses a pitch estimate to determine the fundamental frequency. Experimental results indicate that the exact location of a harmonic peak is critical only in the lower end of the spectrum (less than 2 kHz).

Since the calculation of the spectrum is computationally expensive it is advantageous to downsample the speech by a factor of two. Overlapping Hamming windows are applied to the downsampled speech. The Hamming window has a length of 30 ms centered around the 20 ms analysis frame. The downsampled and windowed speech data are then padded with zeroes to form the sequence of data points on which the Fourier transform is performed. In order to minimize the real-time requirements, a 256-point FFT is computed, yielding a 16 Hz resolution of the spectrum.

The analysis of a voiced frame consists of a sequential search for harmonic peaks within the spectrum. To find the location of the ith harmonic peak, the slope of the low-resolution spectrum at the multiple of the fundamental is determined. The spectrum is searched in the direction of increasing slope until the first spectral peak is located. If a peak is not found before a point halfway to the next harmonic is reached, the multiple of the fundamental frequency is used.

To more accurately determine the location of a harmonic peak within the spectrum, a quadratic interpolation procedure is applied to the low-resolution spectrum. First, the peak in the low-resolution spectrum is found. Next, a quadratic polynomial is fitted through the peak in the low-resolution spectrum S_k and the two adjacent points S_{k-1} and S_{k+1}. The frequency f_i at the peak of the interpolated spectrum is found by

$$f_i = \frac{S_{k-1} - S_{k+1}}{2\,(S_{k-1} - 2S_k + S_{k+1})}. \qquad (1)$$

This interpolation procedure approximates a 2 Hz frequency resolution. The difference between the multiple of the fundamental frequency and the actual location of the peak is defined as a harmonic offset.

As each harmonic peak f_i is found, the pitch estimate for the frame is readjusted. The equation for the ith adjusted pitch estimate p_i is

$$p_i = \frac{\sum_{j=1}^{i} f_j}{\sum_{j=1}^{i} j}, \quad i > 0. \qquad (2)$$

The new fundamental frequency estimate is used to calculate the theoretical frequency of the following harmonic f_{i+1}. This improved estimate corrects for inaccuracies in the original pitch estimate and increases the probability that the spectral peaks will be located within the search regions.

After the harmonic peaks are found, the harmonic offsets are calculated using the adjusted fundamental frequency estimate. The adjustment procedure insures that the offset distribution is symmetric about zero. This symmetry permits efficient coding techniques when preparing the harmonic offset parameters for transmission. Table 1 lists the parameters that are computed, coded, and transmitted for each type of voicing state.

Voiced Frames	Unvoiced Frames	
	Multipulse	Noise
• voiced/unvoiced decision	• voiced/unvoiced decision	• voiced/unvoiced decision
• LPC reflection coefficients	• multipulse/noise decision	• multipulse/noise decision
• speech frame energy	• LPC reflection coefficients	• LPC reflection coefficients
• pitch	• LPC residual frame energy	• LPC residual frame energy
• harmonic frequency offsets	• pulse locations	
	• pulse amplitudes	

Table 1. Coded parameters transmitted for each type of speech frame.

SYNTHESIS

The synthesis portion of the speech coder operates in one of three states depending on the three-level voicing decision. For unvoiced frames, speech is synthesized by using either a random noise or a pulse sequence excitation as input to an LPC lattice filter [7,9]. Efficient computation of the pulse sequence is described in [7]. The harmonic structure of voiced frames is reproduced during synthesis by summing sinusoids at the correct frequencies and amplitudes. Fig. 2 illustrates the major components of the synthesis portion of the speech coder.

The synthesized speech samples are denoted by y_n. The sinusoidal model for the nth synthesized data point in a voiced frame is

$$y_n = \sum_{i=1}^{M} \alpha_{n,i} \sin(\phi_{n,i}), \quad 0 \le n \le N-1, \qquad (3)$$

where i is the harmonic number, M is the total number of harmonics, $\alpha_{n,i}$ is the instantaneous amplitude of the ith harmonic, $\phi_{n,i}$ is the instantaneous phase of the ith harmonic, and N is the number of samples per frame. Instantaneous phase is defined by

$$\phi_{n,i} = \phi_{n-1,i} + 2\pi\omega_{n,i}T, \quad 1 \le i \le M, \qquad (4)$$

where $\omega_{n,i}$ is the instantaneous frequency and T is the sampling period. The initial phase of each harmonic at the onset of voicing is assumed to be zero.

During voiced speech synthesis, the instantaneous harmonic frequencies and amplitudes are obtained using

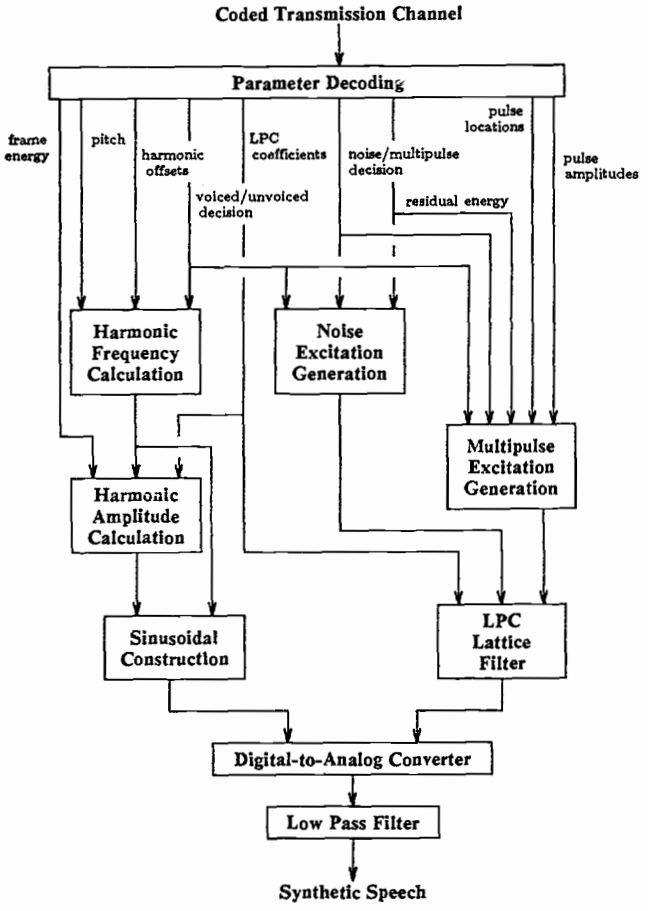

Fig. 2
Synthesizer Block Diagram

linear interpolation on corresponding (i. e., of the same harmonic number) frequency and amplitude information from adjacent frames. This interpolation strategy is valid given the rate of change of the fundamental frequency of human speech. It seems natural to use harmonic amplitudes computed from the FFT for synthesis. However, using harmonic amplitudes computed from an LPC-generated spectral envelope is more practical since there is no need to transmit individual harmonic amplitude information. In practice, there is no significant difference in the resulting synthetic speech.

Changing pitch often causes the number of harmonics in adjacent frames to differ. Since these harmonics are matched sequentially starting at the lowest frequency, harmonics must be added or deleted at the upper end of the spectrum. Harmonics which are added to the present frame are introduced by maintaining a constant frequency over the first half of the frame while linearly increasing the instantaneous amplitude from zero. The initial phase of each harmonic is assumed to be zero. Similarly, harmonics which are deleted are kept at a constant frequency over the second half of the frame while the amplitude linearly decreases to zero.

Energy matching is performed with an orthogonal sinewave approximation. Using Parseval's theorem [10], the energy of the sum of the M sinewaves (assuming constant frequency) is matched to the energy of the input speech signal using the following gain:

$$G = \sqrt{\frac{E}{\frac{1}{2}\sum_{i=1}^{M} \alpha_i^2}} \tag{5}$$

where E is the input speech signal energy.

HARDWARE IMPLEMENTATION

A real-time full-duplex 4.8 kb/s voice coder based on the techniques presented in this paper has been implemented in hardware using two Texas Instruments TMS320C25 Digital Signal Processors. Since the parameters transmitted include those used by a typical LPC vocoder, the system is downward compatible with a 2.4 kb/s LPC vocoder. To operate as an LPC vocoder, the harmonic offsets are not transmitted, and the synthesizer merely substitutes exact multiples of the fundamental frequency for the harmonics.

To achieve an efficicnt hardware implementation, several algorithmic modifications are necessary. For a low-pitched male speaker, the number of harmonic offsets to be transmitted can be large, requiring an excessive data rate. The current hardware implementaiion locates and transmits only the first ten offsets. These are the most important offsets for optimum speech quality. More efficient quantization schemes, such as vector quantization, would allow more offsets to be transmitted, but current real-time limitations prohibit the use of more computationally intensive quantization procedures.

Another consideration is the size of the FFT which can be computed. Since small errors in location of the low-frequency harmonic peaks may introduce large errors when searching for higher frequency harmonics, it is important to compute an FFT with as much resolution as possible. Hardware constraints limit the maximum FFT size to 256 points. The downsampling and interpolation techniques provide an acceptable solution to achieve the required resolution for accurate placement of the harmonics.

The synthesizer is similarly limited by computational constraints. While the harmonic frequency and amplitude interpolators could be implemented using quadratic functions, these are computationally far too expensive. The synthetic speech quality obtained using linear interpolators is similar and the procedure is much more efficient.

CONCLUSIONS

This paper describes a harmonic coding algorithm which may be implemented in relatively modest hardware. This coder, though not of transparent quality, produces synthetic speech which is noticeably smoother and more intelligible than a conventional 2.4 kb/s LPC vocoder. In

addition, the coder does not produce the background distortions during voicing which are characteristic of LPC vocoders. These improvements are most noticeable for male speakers and low pitched female speakers. Given more computational power in the hardware, the performance of the coder can be made to approach that of higher rate harmonic coders [6]. Methods of improving quality beyond that of the system presented here are currently under investigation. A hardware prototype is available for demonstration.

REFERENCES

[1] D. P. Prezas, J. Picone, and D. L. Thomson, "Fast and accurate pitch detection using pattern recognition and adaptive time-domain analysis," *IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, April 1986, pp. 109-112.

[2] T. E. Tremain, "The government standard linear predictive coding algorithm," *Speech Technology*, April 1982, pp. 40-49.

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

[4] I. M. Trancoso and B. S. Atal, "Efficient procedures for finding the optimum innovation in stochastic coders," *IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, April 1986, pp. 2375-2378.

[5] L. B. Almeida and J. M. Tribolet, "A spectral model for nonstationary voiced speech," *IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, May 1982, pp. 1303-1306.

[6] R. J. McAulay and T. F. Quatieri, "Mid-rate coding based on a sinusoidal representation of speech," *IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, March 1985, pp. 945-948.

[7] D. L. Thomson and D. P. Prezas, "Selective modeling of the LPC residual during unvoiced frames: white noise or pulse excitation," *IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, April 1986, pp. 3087-3090.

[8] J. Burg, "A new analysis technique for time series data," *Proc. NATO Advanced Study Institute on Signal Proc.*, Enschede Netherlands, 1968.

[9] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural sounding speech at low bit rates," *IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, May 1982, pp. 614-617.

[10] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Prentice Hall, Inc., Englewood Cliffs, N. J., 1975.