TWO NOVEL SPEECH CODING TECHNIQUES BASED ON MULTIPULSE REPRESENTATION

O. A. Alim, E. A. Youssef and A. G. Mokhtar

Department of Electrical Engineering
Faculty of Engineering
University of Alexandria

ABSTRACT - Two new coding techniques based on multipulse representation of the excitation signal in the time and frequency domains are presented. Algorithms, illustrations, modelling processes as well as bit rates are discussed. Real speech, English as well as arabic mono and di syllabic words were used to test and evaluate the coding techniques subjectively and objectively. Both coding techniques produced good quality speech but the time domain method which operates at bit rates in the range of 8-16 kbps, was found to be superior to the frequency domain method which operates at a bit rate of 5740 bit/s.

INTRODUCTION

One of the techniques which is most promising in the area of speech coding is Linear Predictive Coding (LPC) . This technique when applied, produced synthetic speech quality at a bit rate of 2.4 kbits/s [2]. Slight quality improvements were possible as bit rates were increased above 2.4 kbits/s, but the degradations inherent in the simplistic LPC model remain. The major impediment to toll or communications quality speech lies in the single pulse periodic excitation, which adds a mechanical aspect to the synthetic speech. As a result, it was clear that the key for an improved LPC system performance was excitation improvement. A breakthrough was made by B. S. Atal and J. R. Remede in 1982 [1] when they introduced the multipulse excitation in which they used more than one pulse per pitch period and adjusting the individual pulse positions and amplitudes to minimize a certain error criteria. In 1989 J. V. Schalkwyk and J.V. Der Linde [4] used a frequency domain multipulse coding technique at 2.4 kbits/s promising to have much more natural sounding speech than the conventional LPC system.

Inspired by the work done in the excitation improvement area, two new coding techniques based on multipulse representation of the excitation signal are presented in this paper. The first technique is based on the multipulse excitation in the time domain. The second technique is based on the multipulse excitation in the frequency domain.

TIME DOMAIN METHOD

The algorithm presented here uses the LPC error sequence generated by passing the original speech through the LPC inverse filter, namely

$$e(n) = S(n) - \sum_{k=1}^{p} a_k S(n-k) \qquad (1)$$

where a_k's are the coefficients resulting from LPC analysis and p is the order of the LPC filter. The objective is to obtain another excitation sequence which contains a smaller number of pulses, so when passed through the synthesis filter, the produced speech resembles the original one. Equ (1) when rewritten as

$$S(n) = e(n) + \sum_{k=1}^{p} a_k S(n-k) \qquad (2)$$

shows that higher values of e(n) affect the output speech more than lower ones. Also the energy of the error sequence is given by Equ. (3), with w(n-m) being the window used.

$$E_n = \sum_{-\infty}^{\infty} e^2(m)\, w(n-m) \qquad (3)$$

Plotting E_n along a frame of N samples [figures 1 & 2], it is quite clear and reasonable that the energy is high at the interval of high amplitude error pulses.

Algorithm Formulation

A certain number of pulses, X, are to be used to form the new excitation sequence. Two paths arise, the first one for voiced speech. From the energy waveform, it could be seen that high energy is concentrated at the beginning of the pitch periods, where high amplitude error pulses exist. This is in accordance with the fact that, in the digital speech production model, the excitation for voiced speech consists of a train of impulses separated by a time period equal to the pitch period. From the energy waveform, the algorithm detects the intervals of high energy. Assume that T high energy intervals are detected. From each region, the highest absolute Y pulses are selected. Thus a sequence called the high energy sequence he(n) is formed having a length of YxT. A condition exists that YxT< X. After determining the Y pulses, this sequence is subtracted from the original error sequence to result in a new sequence sb(n), i.e.

$$sb(n) = e(n) - he(n) \qquad (4)$$

with the condition YxT < X, let the R be the residual number of pulses

$$R = X - Y * T \qquad (5)$$

The new sequence sb(n) is segmented into R segments and from each one, the highest absolute amplitude pulse is selected to form a residual sequence, re(n) (not to be confused with the error sequence e(n)). The new excitation sequence is formed by adding he(n) to re(n)

$$exc(n) = he(n) + re(n) \qquad (6)$$

This sequence is used to excite the LPC synthesis filter at the receiver. The first part of the excitation signal, he(n) satisfies the motivation of the method, namely: high amplitude error pulses contribute more to the synthesized speech and from the energy diagram, the energy is higher in the intervals that contain high amplitude error pulses. As a result, all the pulses of he(n) are concentrated at T intervals of time. If this signal excites the synthesis filter, then other parts of the frame will suffer due to lack of excitation. The second part of the excitation signal spreads R pulses along all the time axis in the frame, thus avoiding the strict energy concentration problem and resulting in a better reproduction of the synthesized speech. The value of Y is only limited by speech properties, the number of high energy intervals in a frame is an independent variable, thus Y is determined according to YxT < X. For average human beings, the pitch periods range from 4 ms to 17 ms. Thus in a frame of 20 ms, we expect that at most five high energy intervals will be contained in a frame. For safe design, the worst case should be assumed, where T = 5 ,thus

$$X = 5 * Y + R \qquad (7)$$

Therefore, if X = 20, then Y could be 3 or 4. The second path, which is for unvoiced speech, the energy diagrams show no energy concentrations (figures 3 & 4). Thus the error sequence is segmented into X segments and the absolute highest amplitude pulse is selected from each segment. These highest absolute amplitude pulses form the new excitation sequence. It is to be noted that no voiced/unvoiced detection is used as the algorithm determines the value of T and accordingly if $T \leq 5$ it selects the first path, otherwise it selects the second. As expected, the results show that increasing the number of pulses per frame X, the performance was improved. This is seen in performance curves (figures 5 to 7) which are the Mean Square Error (MSE), the Autocorrelation measures and the LPC distance measure, which is the Log Likelihood Ratio (LLR). This result is also confirmed subjectively. On the other hand, Y = 4 seems the best average choice, however for more

accurate results more research has to be done in this point. Looking at the new excitation signal and at the original error sequence, we can say that the new excitation sequence is a result of a "Non-uniform Sampling" operation on the original error sequence. Figures 8 to 11 show an example.

FREQUENCY DOMAIN METHOD

The motivation for this method came from the frequency domain multipulse representation of the excitation signal developed by J. V. Schalkwyk and J. V. Der Linde [4] in which they represented the spectrum of one pitch period of the error signal by 3 pulses and the spectrum was modeled by straight line interpolation between the 3 pulses. In the algorithm presented here, a different approach was taken in the modeling process, from which an improved, more general model is proposed.
The error signal has several properties , most importantly; is that its bandwidth should not be less than 1.5 kHz for proper recovery of the speech signal. This is verified in [3] and [4].

Algorithm Formulation

In this method, the error sequence is available after filtering the speech with the inverse LPC filter (Equ.1). The spectrum of one pitch period of the error is obtained through a Fast Fourier Transform (FFT) operation. Based on the fact that for proper recovery, a minimum bandwidth of 1.5 kHz of the error sequence is necessary, we are only interested in the range from 0 - 2 kHz for a safe margin. The group of pulses that are to represent the spectrum are obtained by determining the location and amplitudes of all the maxima in the frequency spectrum below 2 kHz. The average of the minima in the frequency spectrum is also determined as it will be used in the modeling process.

Different Models

From figures 12 & 13, which display the spectra of one pitch period of different error signals, they all have a basic feature; the shape. The spectrum between each maximum location and minimum location could be approximated by different interpolating functions. From the examination of various error spectra, the options boiled down to:

1) Trigonometric interpolation
2) Parabolic interpolation

To choose between the two methods, modeling of the spectra was carried out using both methods, and the mean squared error was calculated each case as an evaluation measure for the interpolating functions. The number of spectra model runs used to evaluate the interpolating functions was 50 runs. The results show a mean of 8.16% improvement for the trigonometric interpolation over the parabolic one. The results are shown in figure 14. It is to be noted that in all 50 runs, not once the parabolic interpolation resulted in a better modeling than the trigonometric one.

The Modelling Process

Having determined the maxima of the spectrum and the average of its minima, the modeling process using trigonometric interpolation proceeds as follows: The location of the zeros of the modeled spectrum are assumed to be the midpoints between the maxima locations. Trigonometric interpolation is made between the first minimum and the first maximum followed by interpolation between the first maximum and the second minimum. This process continues until the spectrum is completely modeled. At the receiver, and after modeling the spectrum, from the last point of the modeling process (approximately 2000 Hz) to half the sampling frequency, the spectrum is filled with uncorrelated white noise signal of appropriate amplitude. If this noise is omitted, the synthesized speech suffers a certain lack in the high frequency content. An Inverse Fast Fourier Transform (IFFT) operation is carried out on the modeled spectrum resulting in a recovered continuous excitation signal over one pitch period. It is to be noted that the time domain equivalent of the frequency domain multipulse signal is generated by the repetition of one pitch period for the duration of the frame. This signal is used to excite the LPC synthesis filter at the receiver. Figures 15 to 17 show an example.

THE PROPOSED SPEECH CODERS

Two speech coders are proposed based on the time domain method as well as the frequency domain method. The proposed coders are shown in figures 18 & 19. The bit rates at which these coders operate depend on the number of pulses representing the excitation signal as well as the coding of the parameters of the system as a whole. Speech was sampled at 8 kHz, and a frame length of 20 ms was used. For coding purposes, the systems' parameters were divided into three groups:

1) LPC Filter coefficients: Using 10th order LPC filter and using the reflection coefficients representation, these parameters are coded into 37 bits.

2) Pulse amplitudes: 4 bits are used to code each pulse amplitude with a 7-bit amplitude scale factor.

3) Pulse locations: 4 bits are used to code the pulse locations using a differential coding scheme. For the time domain method, using 16 pulses/frame i.e. $X = 16$ the bit rate would be 8.6 kbits/s, with $X = 20$, the bit rate would be 10.2 kbits/s, with $X = 24$, the Bit rate would be 11.8 kbits/s, with $X = 28$, the bit rate would be 13.4 kbits/s and with $X = 32$, the bit rate would be 15 kbits/s. For the frequency domain method, the number of pulses is not the same each frame as it varies according to the spectrum of the error. For this reason a variable frame length transmission is recommended. At the beginning of each frame a word is sent to the receiver indicating the number of pulses to expect. To determine the bit rate of that system, a histogram was made for the number of pulses output from the analysis of each frame. The histogram shows a mean of 8.8554 pulses with a variance of 1.6847. The histogram is shown in figure 20. According to this figure, the proposed system would operate at 5740 bits/s.

RESULTS

For experimental results, real speech was synthesized by the two methods where, 33 English as well as Arabic, mono and di syllabic words were tested. The results were evaluated Subjectively, objectively and through waveform & spectrum examination. In the subjective tests, the listeners were asked to rank the synthesized speech in one of 4 categories, unacceptable, acceptable, good and excellent. Each category was given a weight of 0, 1, 2 & 3 respectively. The results were then averaged to give a rating out of 3 for each synthesized speech segment. Figures 21 to 23 show a comparison of the time domain representations of an original segment of the speech and the synthesized version.

The time domain method produced good quality speech as judged subjectively (2.103/3, on the average) and objectively and was superior to the frequency domain method which also produced good quality speech (1.8921/3 , on the average) however a little bit metallic.

REFERENCES

[1] Atal, B.S. & Remde, J.R. (1982), *A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates*, ICASSP, IEEE International Conference on Acoustics, Speech & Signal Processing., p. 614-617.

[2] Markel, J.D. & Gray, A.H. (1976), *Linear Prediction of Speech*, Springerverlag, New York.

[3] O'Shaughnessy, D. (1987), *Speech Communications, Human and Machine*, Addison-Wesley Publishing Company.

[4] Schalkwyk, J.V. & Der Linde, J.V. (1989), *Frequency Domain Multipulse Coding at 2400 bits/s*, COMSIG' 89, p. 137-142.

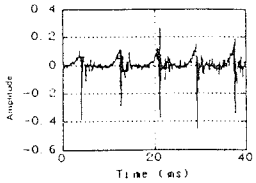


Figure 1. Error signal for voiced speech.

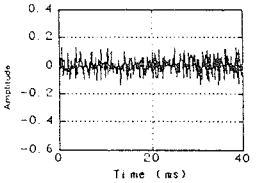


Figure 3. Error signal for unvoiced speech.

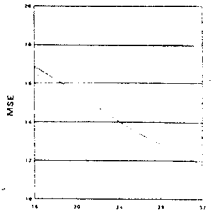


Figure 5. MSE vs. Number of pulses X.

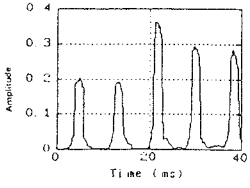


Figure 2. Energy.

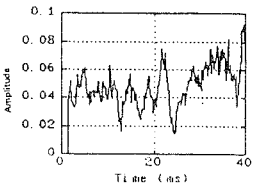


Figure 4. Energy.

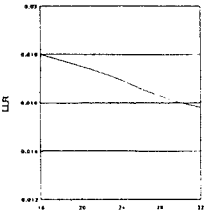


Figure 6. LLR vs. Number of pulses X.

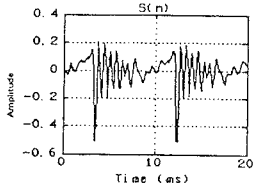


Figure 8. Original speech.

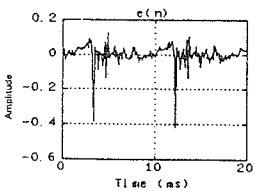


Figure 9. Error Signal.

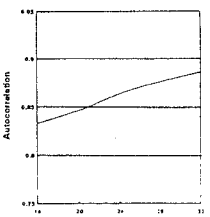


Figure 7. Autocorrelation vs. Number of pulses X.

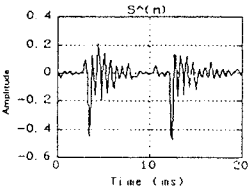


Figure 10. Synthesized speech.

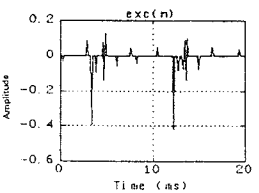


Figure 11. Excitation signal.

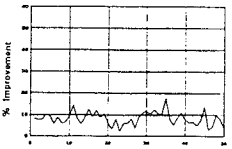
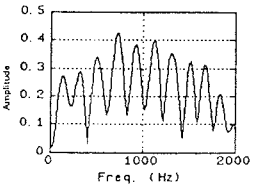


Figure 14. Percentage improvement of trig. interpolation over parabolic one.



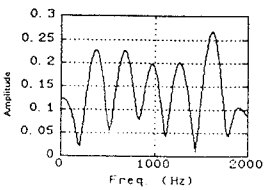


Figure 12 & 13. Error signal spectra.

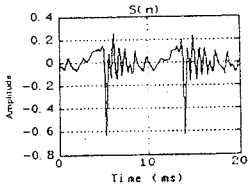


Figure 16. Original speech.

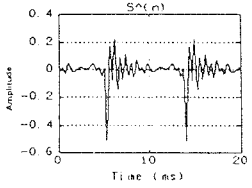


Figure 18. Synthesized speech.

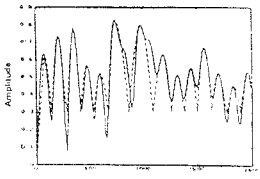
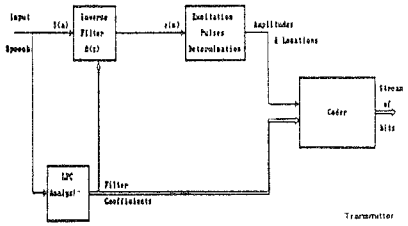
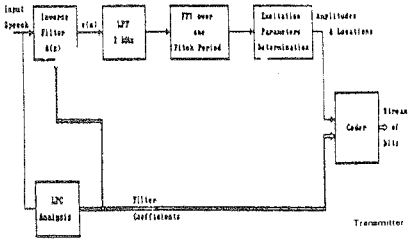


Figure 17. Original & modeled spectra.





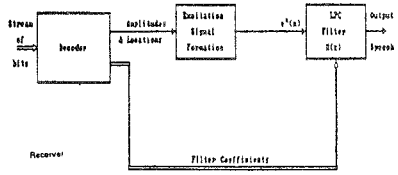


Figure 18. Speech coder based on the time domain method.

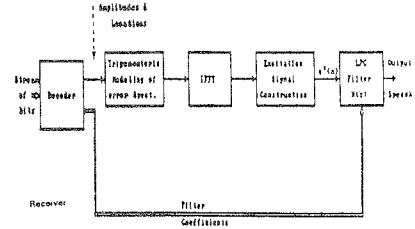


Figure 19. Speech coder based on the frequency domain method.

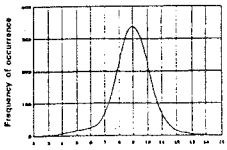


Figure 20. Histogram of the number of pulses.

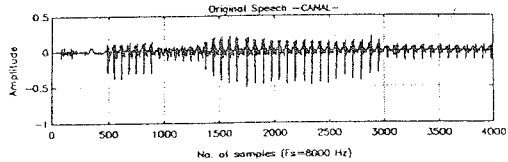


Figure 21.

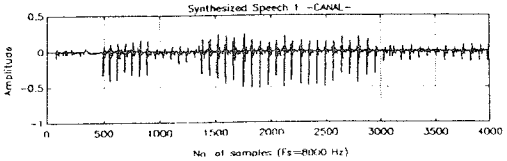


Figure 22.

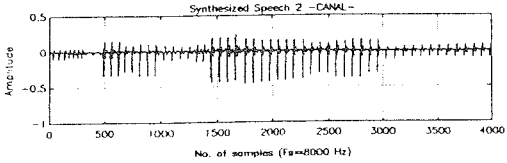


Figure 23.