A NEW MODEL OF LPC EXCITATION FOR PRODUCING NATURAL-SOUNDING SPEECH AT LOW BIT RATES

Bishnu S. Atal and Joel R. Remde

Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

The excitation for LPC speech synthesis usually consists of two separate signals - a delta-function pulse once every pitch period for voiced speech and white noise for unvoiced speech. This manner of representing excitation requires that speech segments be classified accurately into voiced and unvoiced categories and the pitch period of voiced segments be known. It is now well recognized that such a rigid idealization of the vocal excitation is often responsible for the unnatural quality associated with synthesized speech. This paper describes a new approach to the excitation problem that does not require a priori knowledge of either the voiced-unvoiced decision or the pitch period. All classes of sounds are generated by exciting the LPC filter with a sequence of pulses; the amplitudes and locations of the pulses are determined using a non-iterative analysis-by-synthesis procedure. This procedure minimizes a perceptual-distance metric representing subjectively-important differences between the waveforms of the original and the synthetic speech signals. The distance metric takes account of the finite-frequency resolution as well as the differential sensitivity of the human ear to errors in the formant and inter-formant regions of the speech spectrum.

INTRODUCTION

Synthesis of natural-sounding speech at low bit rates has been a topic of considerable interest in speech research. Speech coding methods can be classified into two broad categories: The first type of coders are the *waveform* coders [1] which attempt to mimic the speech waveform as faithfully as possible. The second type of coders are *vocoders* [2-4] which synthesize speech using a parametric model of speech production. Typical examples of waveform coders are pulse code modulation systems and their differential generalizations [1], adaptive predictive [5] and transform coders [6]. The waveform coders are capable of producing high quality speech but only at bit rates above about 16 kbits/sec. The vocoders are efficient at reducing the bit rate to much lower values [7] but do so only at the cost of lower speech quality and intelligibility. The speech quality from vocoders cannot generally be improved by increasing the bit rate.

A model of speech production for synthesizing speech at low bit rates is shown in Fig. 1. The model performs two basic functions. It has a linear filter to model the characteristics of the vocal tract and the spectral shaping of the vocal source. The second part provides the excitation to the linear filter. The model assumes that the speech signal can be classified into one of two classes, voiced and unvoiced, and that the pitch period of the voiced speech segment is known. For voiced speech, the excitation is a quasi-periodic pulse train with delta functions located at pitch period intervals. For unvoiced speech, the excitation is white noise.

This idealized model of speech production is widely used for

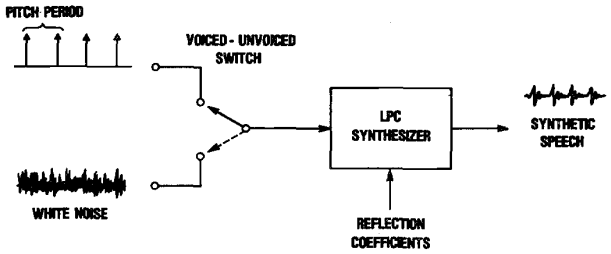

Fig. 1. Block diagram of an LPC speech synthesizer using a periodic pulse train for synthesizing voiced speech and white noise for synthesizing unvoiced speech.

synthesizing speech at low bit rates. However, it is difficult to produce high-quality speech with this model, even at high bit rates. The model is used, because it is perhaps the only way at present to synthesize speech at bit rates below about 4 kbits/s.

Why it is so difficult to produce high-quality speech from this model? The central problem is the highly inflexible way in which the excitation is generated in the model. The introduction of linear predictive coding (LPC) techniques to speech analysis and synthesis did provide an alternate way of representing the spectral information by all-pole filter parameters [8]. However, the model for the excitation of the filter was still carried over from the channel vocoders [3]. The invention of voice-excited vocoders [9] was directed to the solution of the excitation problem but did not provide an entirely satisfactory solution. Accurate separation of speech into two classes, voiced and unvoiced, is difficult to achieve in practice. There are more than two modes in which the vocal tract is excited and often these modes are mixed. An example of a short segment of speech waveform, shown in Fig. 2, illustrates this problem. Of course, there are some easily recognizable voiced and unvoiced portions. But, there are regions where it is not clear whether the signal is voiced or unvoiced and what the pitch period is. It is indeed a difficult task - using either manual or automatic means - to classify short segments of waveform reliably into voiced and unvoiced categories.

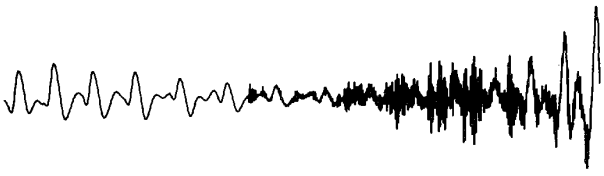

Fig. 2. An example of a short segment of the speech waveform.

Even when the speech waveform is clearly periodic, it is a gross simplification to assume that there is only one point of excitation in the entire pitch period [10]. There is some evidence

CH 1746-7/82/0000 0614 $ 00.75 © 1982 IEEE

that, apart from the main excitation which occurs at glottal closure, there is secondary excitation, not only at glottal opening and during the open phase, but also after the closure [10]. These results suggest that the excitation for voiced speech should consist of several pulses in a pitch period rather than just one at the beginning of the period. The main difficulty so far in using a multi-pulse model has been our inability to develop a satisfactory procedure for determining these pulses.

We present in this paper a multi-pulse model for the excitation of LPC synthesizers and describe a simple procedure for determining the locations and amplitudes of the pulses. We find that this model is flexible enough to provide high quality speech even at low bit rates. We make no a priori assumption about the nature of the excitation signal. The excitation consists of a sequence of pulses for all speech classes - including voiced and unvoiced speech. We make no attempt to make it periodic or non-periodic.

Of course, if the number of pulses is increased to arbitrarily large value so that there is a pulse at every sampling instant, it should be possible to duplicate the original speech waveform (at the expense of a high bit rate). What we find interesting is that only a few pulses (typically 8 pulses every 10 msec) are sufficient for generating different kinds of speech sounds, including voiced and unvoiced, with little audible distortion. Thus, we do not need any prior knowledge of either the voiced-unvoiced decision or the pitch period for synthesizing speech. The periodic (or non-periodic) nature of the speech signal does not play a crucial role in analysis or synthesis, although the periodicity could be used to decrease the bit rate to lower values.

MULTI-PULSE EXCITATION MODEL

Fig. 3 shows the block diagram of an LPC speech synthesizer with multi-pulse excitation. It is quite similar to the traditional LPC synthesizer except for the absence of the pulse and white noise generators and the voiced-unvoiced switch. The excitation for the all-pole filter is generated by an excitation generator that produces a sequence of pulses located at times $t_1, t_2, \cdots, t_n, \cdots$ with amplitudes $\alpha_1, \alpha_2, \cdots, \alpha_n, \cdots$, respectively. The all-pole filter could be replaced by a pole-zero filter if desired. The sampled output of the all-pole filter is passed through a low-pass filter to produce a continuous speech waveform $\hat{s}(t)$. We will now discuss an analysis-by-synthesis procedure for determining the locations and amplitudes of the excitation pulses.

Analysis-by-Synthesis Method of Determining Excitation

An analysis-by-synthesis procedure for determining the locations and the amplitudes of the pulses is shown in the block diagram of Fig. 4. The LPC synthesizer produces samples $\hat{s}_n$ of synthetic speech signal in response to the excitation v_n. The synthetic speech samples are compared with the corresponding speech samples of the original (natural) speech signal to produce an error signal e_n. This error is not very meaningful and must be modified to take account of how the human perception treats the error. Our knowledge in this area is not perfect but phenomena like masking and the limited frequency resolution of the ear must be considered [11,12]. Broadly speaking, the error in the formant regions must be de-emphasized. This is done by a linear filter which suppresses the energy in the error signal in the formant regions. The error signal is thus weighted to produce a subjectively-meaningful measure of the difference between the signals $\hat{s}_n$ and s_n. The weighted error is squared and averaged over a short time interval 5 to 10 msec in duration to produce the mean-squared weighted error ϵ. The locations and amplitudes of the pulses are chosen to minimize the error ϵ.

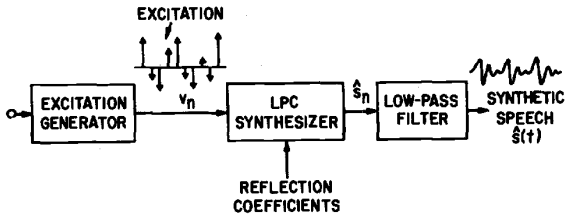

Fig. 3. Block diagram of an LPC speech synthesizer with multi-pulse excitation.

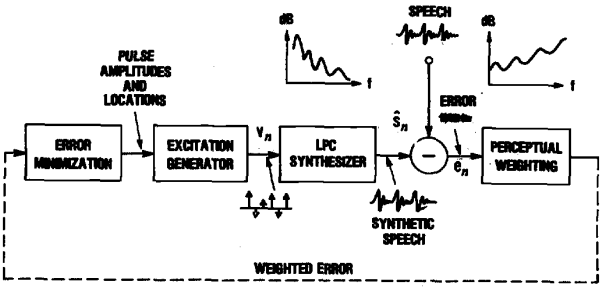

Fig. 4. Block diagram of an analysis-by-synthesis procedure for determining locations and amplitudes of pulses for the multi-pulse excitation.

Perceptual Criteria for Selecting the Error-Weighting Filter

The error-weighting procedure needs further explanation. As suggested earlier, the error e_n is not a valid measure of the perceptual difference between the original and the synthetic speech signals. To obtain a better measure of this difference, we define a frequency-weighted error as

$$\epsilon = \int_0^{f_s} |S(f) - \hat{S}(f)|^2 \, W(f) \, df, \tag{1}$$

where $S(f)$ and $\hat{S}(f)$ are the Fourier transforms of the original and the synthetic speech signals, respectively, $W(f)$ is a suitably-chosen weighting function, f is the frequency variable, and f_s is the sampling frequency. Due to the relatively high concentration of speech energy in formant regions, we can tolerate larger errors in the formant regions in comparison to the in-between formant regions. The weighting function $W(f)$ is chosen to de-emphasize formant regions in the error spectrum. Let $1-P(z)$ be the LPC inverse filter in the z-transform notations. Then, the short-time spectral envelope of the speech signal is given by

$$S_e(f) = |K/[1 - P(e^{-2j\pi f/f_s})]|^2, \tag{2}$$

where K is the mean-squared prediction error. The inverse filter $1-P(z)$ is given in terms of the inverse filter coefficients a_k as

$$1 - P(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} = 1 - \sum_{k=1}^{p} a_k e^{-2j\pi fk/f_s}. \tag{3}$$

If we now represent the transfer function of the error-weighting filter by $W(z)$, then a suitable choice for $W(z)$ is given by

$$W(z) = [1 - \sum_{k=1}^{p} a_k z^{-k}]/[1 - \sum_{k=1}^{p} a_k \gamma^k z^{-k}], \tag{4}$$

where γ is a fraction between 0 and 1 and controls the increase in the error in the formant regions. The filter $W(z)$ changes from $W(z)=1$ for $\gamma=1$ to $W(z)=1-P(z)$ for $\gamma=0$. The value of γ is determined by the degree to which one wishes to de-emphasize

the formant regions in the error spectrum. The optimum value of γ must be determined by suitable listening tests. We find that the choice of γ is not too critical - a typical value is approximately 0.8 at a sampling frequency of 8 kHz. Figure 5 shows an example of the speech spectrum and the frequency response of the corresponding error-weighting filter.

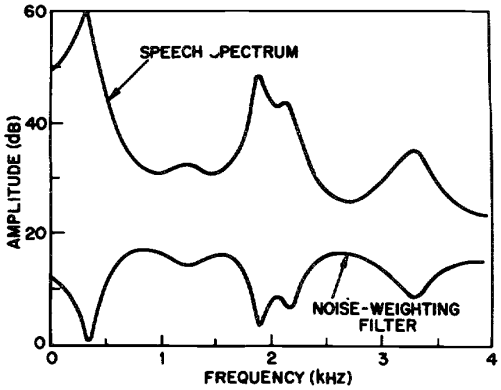

Fig. 5. An example of the speech spectrum and the frequency response of the corresponding error-weighting filter.

Error-Minimization Procedure

The weighted error is used in an error minimization procedure to select the optimal locations and amplitudes of the pulses in the excitation signal. In general, such a procedure would be extremely complex if one seeks to determine all the pulses at once. We find that an efficient solution is obtained by determining the location and amplitude of pulses - one pulse at a time. No doubt, it is a sub-optimal procedure, but our experience so far suggests that it is a reasonable one.

In finding the locations and amplitudes of the pulses - one pulse at a time - we have converted a problem with many unknowns to just two unknowns. Moreover, the amplitude of the pulses appears as a linear factor in the error and as a quadratic factor in the mean-squared error. A closed-form solution for the pulse amplitude is obtained by setting the derivative of the mean-squared error with respect to the unknown amplitude to zero. The mean-squared error is then a function of only the location of the pulse. The optimum location is found by computing the error for all possible pulse locations in a given time interval and by locating the minimum of the error. The procedure can be further refined by observing that the unknown amplitudes of all the pulses can be determined in a single step by solving a set of linear equations provided locations of the pulses are known. We determine the pulse locations first, one pulse at a time, and then use the resulting locations to determine the amplitudes in a single step.

The procedure for finding the locations and amplitudes of various pulses in any given time interval can be summarized as follows: At the beginning, without any excitation pulse, the synthetic speech output is entirely generated from the memory of the all-pole filter from previous synthesis intervals. The contribution of this past memory is subtracted out from the speech signal and the location and amplitude of one single pulse, which minimizes the mean-squared weighted error, is determined. A new error signal is now computed by subtracting out the contribution of the pulse just determined. The process of locating new pulses to reduce the mean-squared weighted error is continued until the error is reduced to acceptable limits. Figure 6 illustrates the error minimization procedure. The waveform in the first row of Fig. 6 is that of the original speech signal, about 5 msec in duration. At the beginning, the excitation is zero and the synthetic speech signal is produced by the memory hangover from the previous synthesis intervals. The error of course is large and is reduced by placing a pulse in the excitation as shown in Fig. 6 (b). This process is continued by adding more pulses; the results with two pulses are shown in Fig. 6 (c). The error is decreased further. The rate of improvement usually slows down after a few pulses have been placed. The results with three and four pulses are also shown in Fig. 6. The process of adding pulses can be continued but in practice we find that only minimal improvements are achieved after about 8 pulses have been placed in a 10 msec interval.

RESULTS

Several examples of the original and the synthetic speech waveforms are shown in Figs. 7-9. The duration of the speech segment is 0.1 sec in each figure. An all-pole filter with 16 poles was used to synthesize speech. The parameters of the all-pole filter were determined once every 10 msec using the stablized covariance method of LPC analysis with high-frequency correction [5]. The covariance matrix was computed by averaging speech data over 20 msec long time intervals. The pulse locations and amplitudes were determined by minimizing the error over successive 5 msec time intervals.

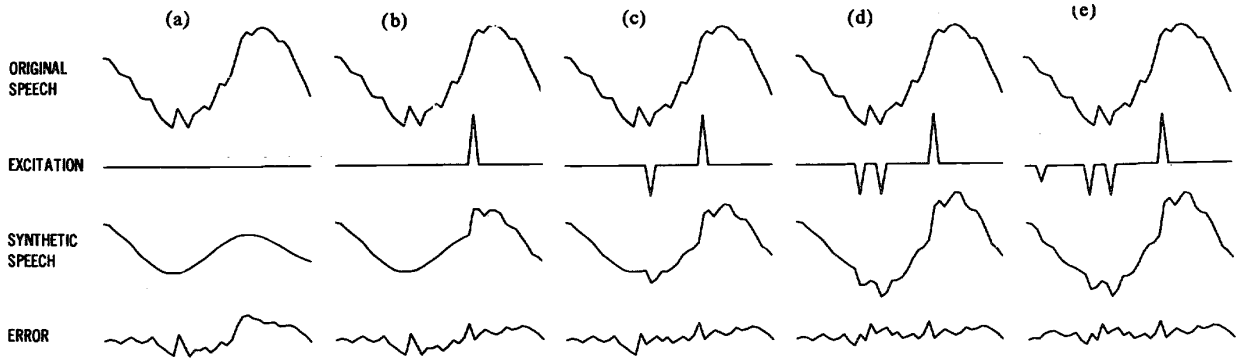

Fig. 6. Illustration of the waveform matching procedure. The waveforms of the original speech, the excitation, the synthetic speech, and the error signals are shown (a) in the beginning without any excitation pulse, (b) with one excitation pulse, (c) with two excitation pulses, (d) with three excitation pulses, and (e) with four excitation pulses.

The figures show the waveforms of the original speech signal, the synthetic speech signal, the excitation signal, and the error between the original and the synthetic waveforms. As can be seen, the multi-pulse excitation is able to follow rapid changes in speech waveforms, such as those occuring in rapid transitions. For voiced speech, the quasi-periodic nature of the excitation is produced by the error-minimization procedure without any knowledge of whether the segment is voiced or unvoiced and what its pitch period is. Similarly, the random excitation pattern is also produced automatically by the matching procedure. Mixed excitation comes out naturally depending upon the nature of the speech signal.

Informal listening tests show that the synthetic speech signal is perceptually close to the original speech signal. The synthetic speech signal from the multi-pulse excitation model does not have the unnatural or buzzy characteristics so often associated with synthetic speech from vocoders.

CONCLUSIONS

Our work on developing a new model of LPC excitation for producing high quality speech is as yet very preliminary. We find that speech quality can be significantly improved by using multi-pulse excitation signal. Moreover, the locations and amplitudes of the pulses in the multi-pulse excitation can be determined by using a computationally efficient non-iterative analysis-by-synthesis procedure that can minimize the perceptual difference between the natural and the synthetic speech waveforms. The difficult problems of voiced-unvoiced decision and pitch analysis are eliminated.

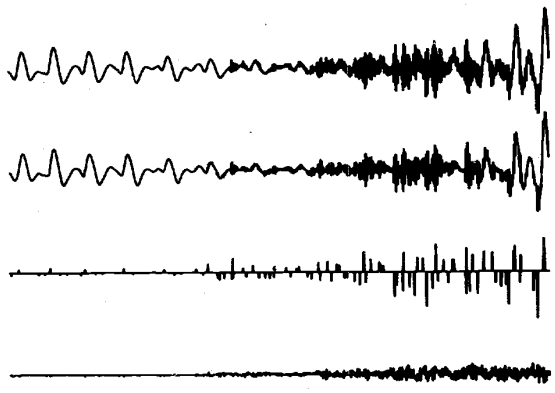

Fig. 7. Waveforms of the original speech, the synthetic speech, the excitation, and the error signals.

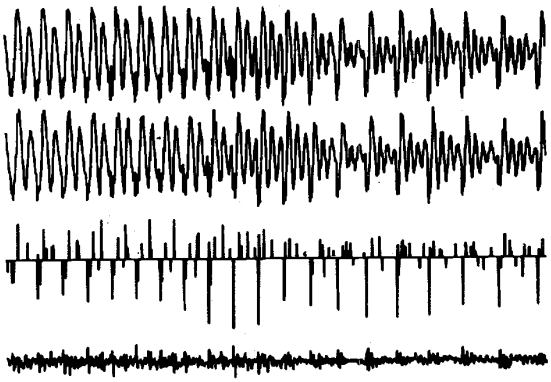

Fig. 8. Another example of the waveforms of the original speech, the synthetic speech, the excitation signal, and the error.

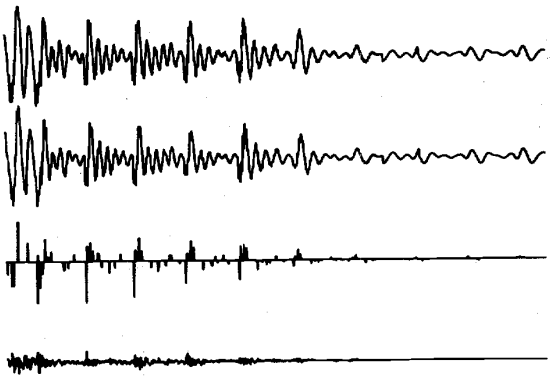

Fig. 9. Another example of the waveforms of the original speech, the synthetic speech, the excitation signal, and the error.

REFERENCES

[1] N. S. Jayant, "Digital Coding of Speech Waveforms: PCM, DPCM and DM Quantizers," *Proc. IEEE*, vol. 62, pp. 611-632, May 1974.

[2] M. R. Schroeder, "Vocoders: Analysis and Synthesis of Speech," *Proc. IEEE*, vol. 54, pp. 720-734, 1966.

[3] J. L. Flanagan, *Speech Analysis Synthesis and Perception*, 2nd Edition, New York: Springer-Verlag, 1972.

[4] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*, New York: Springer-Verlag, 1976.

[5] B. S. Atal and M. R. Schroeder, "Predictive Coding of Speech Signals and Subjective Error Criteria," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 247-254, June 1979.

[6] R. E. Crochiere and J. M. Tribolet, "Frequency Domain Coding of Speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 512-530, Oct. 1979.

[7] J. D. Markel and A. H. Gray, Jr., "A Linear Prediction Vocoder Simulation Based Upon the Autocorrelation Method," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 124-134, 1974.

[8] B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction," *J. Acoust. Soc. Amer.* vol. 50, pp. 637-655, Aug. 1971.

[9] M. R. Schroeder, "Recent Progress in Speech Coding at Bell Telephone Laboratories," *Proc. III Int. Congress on Acoustics*, pp. 201-210, Elsevier Publishing Co., Amsterdam, 1961.

[10] J. N. Holmes, "Formant Excitation Before and After Glottal Closure," *Conf. Rec. 1976 IEEE Int. Conf. Acoust., Speech and Signal Processing*, pp. 39-42, April 1976.

[11] M. R. Schroeder, B. S. Atal and J. L. Hall, "Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear," *Jour. Acoust. Soc. Amer.*, vol. 66, pp. 1647-1652, Dec. 1979.

[12] M. R. Schroeder, B. S. Atal and J. L. Hall, "Objective Measure of Certain Speech Signal Degradations Based on Properties of Human Auditory Perception," in *Frontiers of Speech Communication Research* edited by B. Lindblom and S. Ohman, London: Academic Press, 1979, pp. 217-229.