Proceedings of The Institute of Acoustics

THE USE OF MULTI-PULSE EXCITATION IN SPEECH SYNTHESIS FROM TEXT

A.P.VARGA AND F.FALLSIDE
CAMBRIDGE UNIVERSITY, ENGINEERING DEPARTMENT.

Introduction

A new method of vocoding has recently been introduced by Atal and Remde [1]. This takes the viewpoint that several of the deficiencies of the Linear Prediction (L.P.) method can be corrected by retaining the L.P. filter but modifying its excitation to a "multi-pulse excitation", typically 8 pulses every 10 mS. It makes no distinction between voiced and unvoiced speech. The method results in a significant improvement in speech quality compared with conventional L.P. speech, although considerably more processing is required.

This paper addresses, in a preliminary way, the question of whether the method might provide improved quality in speech synthesis systems. Speech systems employing conventional L.P. suffer from its disadvantages, there are for example problems with unvoiced sounds, pitch extraction and voiced/unvoiced decision making. Also conventional synthesis generally takes no account of the micro-variations in pitch period and gain which exist in natural speech. Multi-pulse excitation mitigates these problems. However the rules used by text-to-speech systems always specify the synthesiser excitation in terms of pitch in the case of voiced sounds, noise amplitude for unvoiced sounds, or a combination of both. It is therefore necessary to relate these rules to multi-pulse excitation if possible. This paper reports on some early investigations into these relationships.

Multi-Pulse Excitation Analysis Method

The analysis is an analysis-by-synthesis technique, shown in fig 1. Whereby for the current frame of speech the L.P. coefficients are first calculated by the covariance method, and then held constant during the subsequent analysis of that frame. The position and amplitude of each pulse of the multi-pulse excitation, for that frame, is found by an iterative procedure. This procedure minimises a perceptually weighted error measure between the spectrum of the original speech frame, S(f), and the spectrum of the current estimate, $\hat{S}(f)$. In continuous form the error measure, ϵ, is:

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

In this W(f) is the spectrum of the perceptual weighting function and f_s is the sampling frequency. It can be shown that a suitable weighting funtion is:

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

where a_k is the kth prediction coefficient, and γ is a fraction which determines the degree of de-emphasis of the formant errors [1].

To start with, a single pulse is applied and ϵ is calculated with this pulse at each position in the frame, in turn. The position which gives the minimum ϵ is then selected and the amplitude of the pulse is found by differentiation. The effect of this pulse is retained and the procedure repeated to find the next position and amplitude. This procedure can be repeated as required; Atal and Remde found that 8 pulses every 10 mS were sufficient.

For the results given here, an analysis frame of 64 samples was used, speech was sampled at 10KHz and the L.P. frame was 256 samples updated every 128 samples.

Characteristics of Multi-Pulse Excitation

The quality of the synthesised speech is dependent on the speaker. A male voice can be modelled well with four pulses per 6.4 mS, while a female voice needs about seven, due to the higher female pitch. Some sample waveforms for a male vowel /a/ at different pitches are shown in figs 2 and 3, and a fricative /ʃ/ in fig 4. Voiced sounds can be matched very well, compare figs 2 a and b; the excitation is shown in part e. For a 4-pulse input, fig 2e, it can be seen that each pitch period has a primary pulse together with some secondary pulses. The primary pulses are associated with the position of maximum prediction error; compare figs 2 e and c. However, the secondary pulses are not so well related to the inverse filter output. Using only a single pulse per pitch period the speech quality is better than conventional L.P. speech, because it allows for micro-variations in the excitation.

The 4-pulse excitation is seen to be a simplified version of the 64-pulse version of fig 2d. The generated excitation is dependent on the relative positions of the frame and pitch period, but this has no effect on the resulting synthesised speech. Compare figs 2 e and f, where in f the analysis frame has been shifted by 22 samples. The effect on the primary pulses is generally small, while the secondary pulses are affected to a much greater degree. The phase relationship results in a beating effect on the positions and amplitudes of the pulses.

Fricatives produce a fairly even pulse distribution. The waveform of the synthesised fricative does not match the original nearly as well as for voiced sounds, see fig 4. Nevertheless in connected speech there is little perceptible degradation. While for sustained fricatives only a very slight "interference" can be heard.

Detailed examination of the waveforms shows that the low frequency matching is always very good, whereas the high frequency match is poorer and sample dependent. This is especially evident when one compares the original and synthesised waveforms of figs 3 a and b. To a degree, this is due to the particular implementation used here. It is believed improvements can be made in several areas, such as the time sampling of the Fourier transforms used to produce $\hat{S}(f)$, and possibly by improved calculation of the prediction coefficients by a method such as that described by Atal and Schroeder [2]. There is only a little perceptible degradation in quality from a sixteenth order filter to a twelfth order. Below twelve the degradation is more noticeable. Also merely adding a few more pulses does not necessarily improve the quality, though, by using a large number of pulses, upwards of 30, the match is almost perfect. The analysis is C.P.U. intensive; ten minutes on a P.D.P. 11/60 for one second of speech, to find four pulses per 6.4 mS frame. However no attempt at optimisation has been made other than to try every second pulse position, instead of every position. There is no significant degradation in quality for every second position. As one increases the step beyond three positions the degradation becomes more significant.

Application To Pitch Alteration By Rule

Attempts have been made to relate multi-pulse excitation to the rules used for intonation control in conventional, single-pulse, text-to-speech synthesis systems. The aim being to find a controllable form of multi-pulse excitaion. At present there is too much change in the amplitude and positions of the pulses from frame to frame for this to be achievable. A number of attempts have been made to vary the pitch by adjusting the timing of the multi-pulse excitation, but these have not been successful, possibly because of the variablity of the pulses from frame to frame.

Conclusion

Multi-pulse excitation used in vocoding can produce high quality speech at low bit rates. While the analysis time described in this paper can be reduced by careful programming, real-time analysis is probably not feasible with the present algorithms and conventional processors. It has not been possible to apply the technique, in its present form, to a simple pitch change by rule system. However, the improvement in quality obtainable justifies further examination, particularly in the area of pitch synchronous analysis. Also further investigation into the manipulation of the pulses is required. The results presented here are the initial findings of a continuing project.

References

1 B.S.ATAL and J.R.REMDE 1982 IEEE proc ICASSP 82, 614-617. A new model of L.P.C excitation for producing natural sounding speech at low bit rates.

2 B.S.ATAL and M.R.SCHROEDER 1979 IEEE trans ASSP 27, 247-254. Predictive coding of speech signals and subjective error criteria.

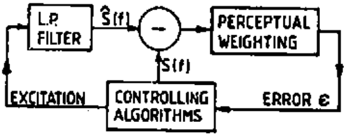

fig 1
Simplfied block diagram of Multi-pulse analysis

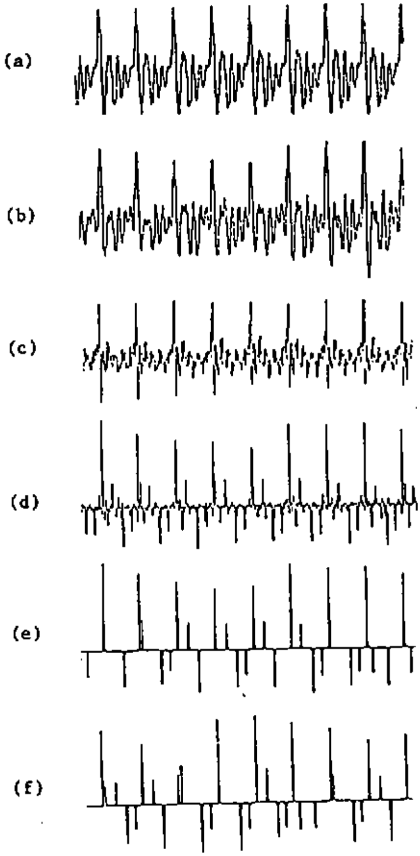

Fig.2, Male vowel /a/

a) original
b) synthesised,4-pulse excitation
c) output of inverse filter
d) 64-pulse excitation
e) 4-pulse excitation
f) 4-pulse excitation for analysis frame offset by 22 samples from that of (e)

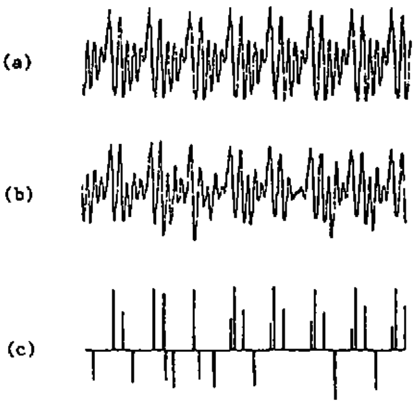

Fig.3, Male vowel /a/, higher pitch than fig.2.

a) original
b) synthesised,4-pulse excitation
c) 4-pulse excitation

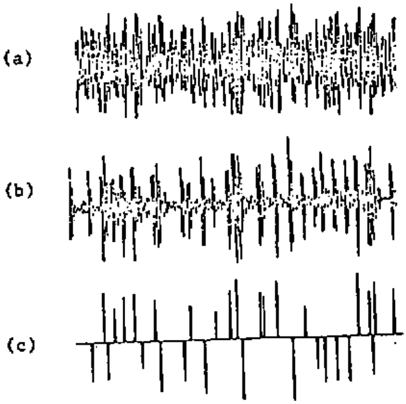

Fig.4, Fricative /ʃ/

a) original
b) synthesised,4-pulse excitation
c) 4-pulse excitation