BRITISH ACOUSTICAL SOCIETY

"SPRING MEETING" at Chelsea College, London, S.W.3 on
Wednesday 25th April / Friday 27th April, 1973.

SPEECH AND HEARING: Session 'B': Speech Analysis and Transmission.

Paper No:
73SHB7

Speech Analysis-Synthesis on a Small Computer

R.G. Crichton and F. Fallside
Cambridge University Engineering Department
Trumpington Street, Cambridge CB2 1PZ

A speech analysis-synthesis system using linear prediction is described. The system has been simulated on a PDP 8I computer with software floating point and runs in approximately 800 x real time. The speech production model is presented and the analysis-synthesis scheme is briefly outlined. The model is related to a low data rate model suitable for articulatory synthesis, and fixed point operation is discussed.

1. Speech Production Model

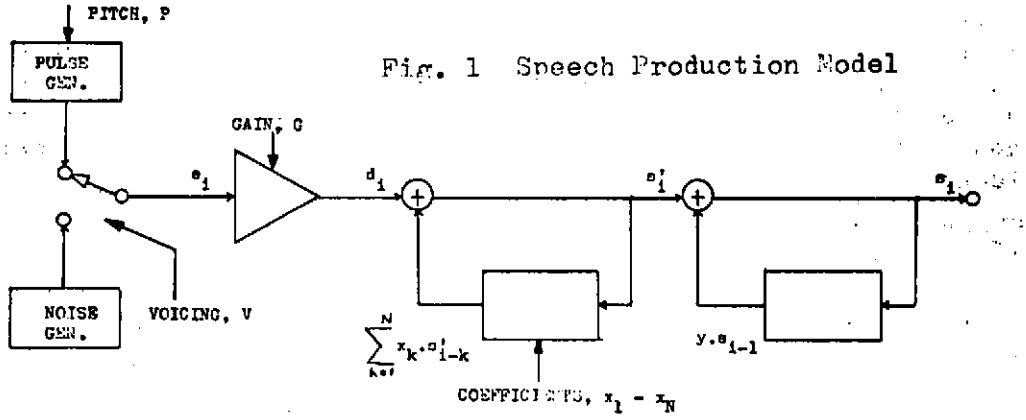

Fig. 1 Speech Production Model

The speech production model is shown schematically in figure 1, being similar to that described by Atal and Hanauer (1). The fixed single pole filter at the output removes the bulk effect of the source and radiation and all processing is carried out on the sequence s'_i.

The Z-plane transfer function of the model is

$$H(z) = \frac{S'(z)}{E(z)} = \frac{G}{1 - \sum_{k=1}^{N} x_k \cdot z^{-k}} \qquad 1$$

The adequacy of this model and the choice of N, the number of coefficients and M, the number of samples per analysis frame are discussed elsewhere (1,2).

2. Analysis

Two similar methods of coefficient extraction have been described, generally referred to as the "autocorrelation matrix" and "covariance matrix" methods. Both involve a least squares minimisation of the input to the filter, d_i, leading to a set of N simultaneous equations in $x_1 - x_N$. The optimum methods of solution differ, the autocorrelation matrix being a special case of the covariance matrix with a rapid, non-iterative solution (5). The

covariance matrix is best solved by the square root method (6). The autocorrelation matrix is used in the present simulation. The two methods are comprehensively compared in reference 3.

The analysis is frame by frame with no overlap. If M is the number of samples per frame,

$$\sum_{k=1}^{N} x_k . r_{|j-k|} = r_j \qquad 2$$

where

$$r_j = \sum_{i=j+1}^{M} s'_{i-j} . s'_i \qquad 3$$

and

$$s'_i = (s_i - s_{i-1}) . W_i \qquad 4$$

where s_i is the input speech sequence, s'_i is the pre-emphasised sequence and W_i is a window function (2). The inverse filter concept (2,4) is used to extract the remaining parameters, i.e. pitch, gain and voicing. The inverse filter output sequence d_i is given by

$$d_i = s'_i - \sum_{k=1}^{N} x_k . s'_{i-k} \qquad 5$$

This is approximated in the synthesis by either a pulse train or a random number sequence. The autocorrelation sequence A_j of d_i shows a peak at the pitch interval if s_i is voiced (see figure 2).

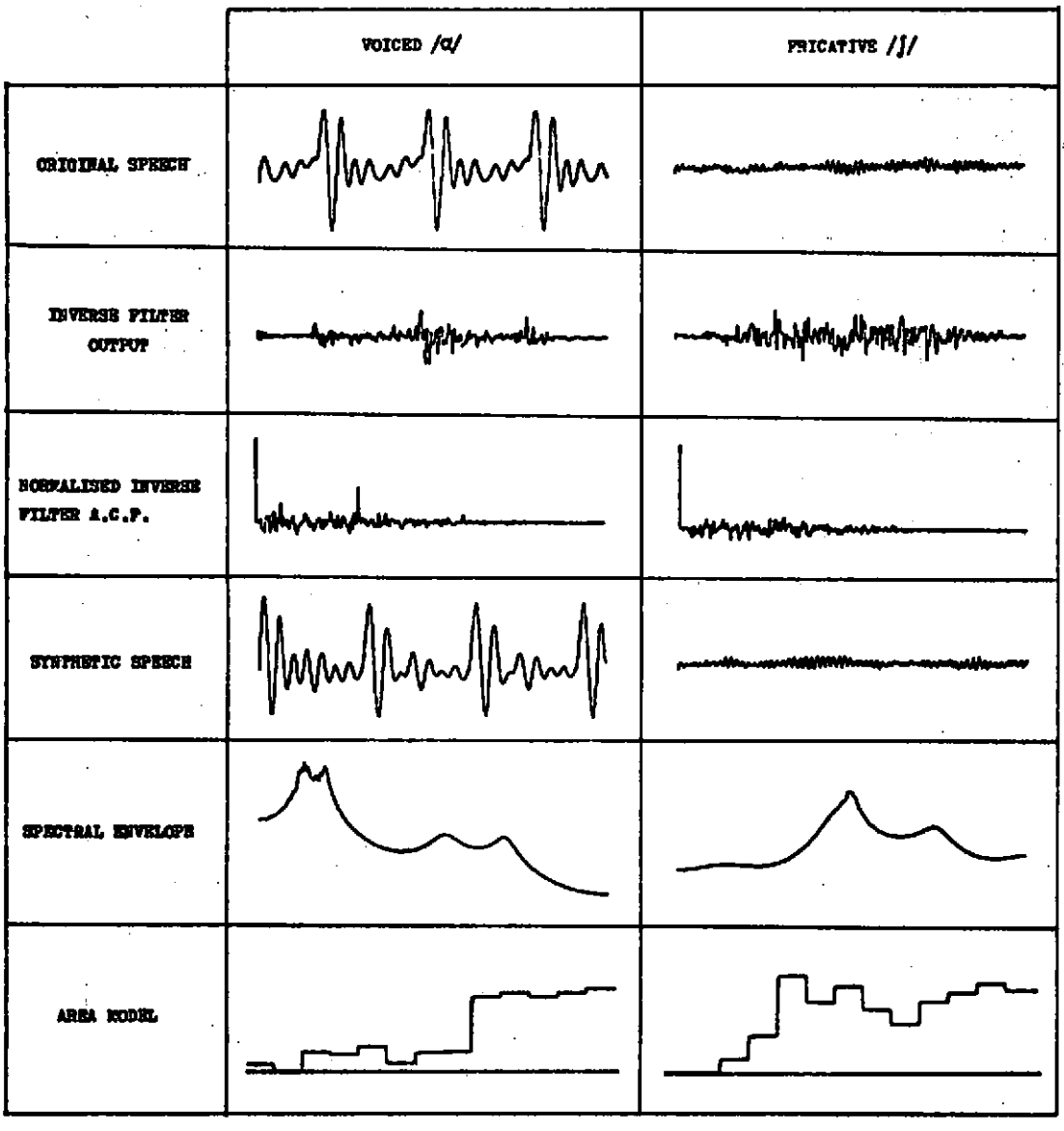

Fig. 2 Analysis and Synthesis of Single Frames

This sequence can be normalised to A_0 when a simple threshold test allows extraction of voicing and pitch (4). Gain is calculated from the total eneregy at the output of the inverse filter (3).

3. Area Representation of Coefficients

A set of reflection coefficients for a vocal tract area model occurs as an intermediate result in the solution of equation 2. The model consists of N abutting, smooth cylinders terminated in a unit acoustic resistance and excited by a volume velocity source (1). These areas have been found to fairly represent the articulatory behaviour of the vocal tract, and it is expected that they could form the final stage of an articulatory model.

4. Spectral Envelope and Formants

The linear prediction model has been widely used to calculate spectral envelopes and track formants (2,3). The spectral envelope of the filter transfer function is

$$H(w) = \frac{1}{1 - \sum_{k=1}^{N} x_k . e^{-jwk}} \qquad 6$$

A Fast Fourier Transform on the impulse response of the inverse filter gives an efficient means of forming H(w) (2). Formants are located either as peaks in H(w) or as roots of the polynomial:

$$1 - \sum_{k=1}^{N} x_k . z^{-k} = 0 \qquad 7$$

Bandwidth and frequency criteria allow formant labelling.

5. Computing Details

The computer is a PDP 8I with 12k x 12 bit core, 64k disc, ADC, DACs, display and plotter. The processor uses hardware unsigned multiply and divide. A flexible and comprehensive program has been written which allows both single frame and continuous analysis and synthesis. Table 1 shows the options available under keyboard control. Double buffering allows 12 bit speech samples to be transferred to and from disc at 10 kHz. The current program

Request N,M and window parameters	
Accept M samples from microphone Extract filter coefficients and areas Form output of inverse filter Form A.C.F. of inverse filter output Extract pitch, gain and voicing Synthesise M samples of speech Analyse and Synthesise M samples Display/plot/punch M length sequence Display/plot areas Form spectrum of filter transfer function Display/plot spectrum Punch all model parameters Accept M length input sequence from tape	Single frame options
Accept continuous speech from microphone Analyse and synthesise continuous speech Play back original or synthetic speech Edit/display/plot continuous speech	Continuous options
Print list of options Return to operating system environment	

Table 1 Keyboard Options for Analysis-Synthesis Program

accommodates 1.6 seconds of both original and synthetic speech which is stored on one 32k disc while the source form of the program resides on the other disc. A fixed point F.F.T. routine (7) allows fairly rapid calculation of the inverse filter output A.C.F. Figure 2 shows some plots generated by the program.

The quality of the synthetic speech is fair for M = 256, N = 12. Experiments with parameter filtering and interpolation, and predictive algorithms for pitch and voicing have indicated that excellent quality speech is feasible.

The computation time is tediously slow for continuous analysis-synthesis. Most of the arithmetic is well bounded and fixed point operation on a modern machine should reduce the analysis-synthesis time to 20 x real time. Dedicated hardware and pruning of the algorithms would bring real time operation within sight.

6. Conclusions

An analysis-synthesis system using linear prediction has been successfully simulated on a small computer. The method has been shown to be extremely versatile, with low data rate capabilities and a simple relationship to the more familiar formant concept.

1. Atal, B.S. and Hanauer, S.L., Speech Analysis and Synthesis by Linear Prediction of the Speech Wave. Journ.Acous.Soc.Am., August 1971.
2. Markel, J.D., Digital Inverse Filtering - A New Tool for Formant Trajectory Estimation. I.E.E.E. Trans. Audio and Electroacoustics, June 1972.
3. Makhoul, J.I. and Wolf, J.J., Linear Prediction and the Spectral Analysis of Speech. Bolt Beranek and Newman, Report no.2304.
4. Markel, J.D., Automatic Formant and Fundamental Frequency Extraction from a Digital Inverse Filter Formulation. 1972 Conf. Speech Communication and Processing, paper B9.
5. Levinson, N., The Wiener R.M.S. Criterion in Filter Design and Prediction. In Wiener, N., Extrapolation and Smoothing of Stationary Time Series, M.I.T. Press, 1966.
6. Fadeev, D.K. and Fadeeva, V.N., Computational Methods of Linear Algebra, W.H. Freeman, 1963.
7. Rothman, J.E., A Fast Fourier Transform Subroutine for Complex Data, DECUS Program Library, no.8-144.