ANALYTIC MODELLING OF SPEECH SPECTRA
FJ OWENS & R LINGGARD
DEPARTMENTS OF ELECTRICAL & ELECTRONIC ENGINEERING
THE ULSTER POLYTECHNIC
THE QUEEN'S UNIVERSITY OF BELFAST

1. INTRODUCTION Previous work in the field of speech resynthesis (analysis/synthesis) have adopted either a time domain or a frequency domain approach. A popular time domain method is LPC (1) which uses an all-pole model in its simplest analytic form. The frequency domain approaches which have enjoyed the most success (2, 3) use iterative techniques.

This paper describes a method whereby the transfer-function of a pole-zero model of the vocal tract may be derived analytically from the short-time amplitude spectrum and how it is applied in the resynthesis of voiced speech, using serial and parallel filter realisations of the model, implemented on a programmable digital speech synthesiser.

2. CURVE-FITTING TECHNIQUE Consider the smoothed spectrum of a typical vowel sound - Fig 1. For the resynthesis of this vowel using any spectral matching technique, the aim is to derive a model with a frequency response which "fits" as closely as possible to the spectrum of the original vowel. It is theoretically possible to derive a model with a frequency response that fits the original spectrum exactly at each of its computed frequency points. However, the high order of such a model would make resynthesis complex and would render the required data rate for the synthesiser unacceptably high. It is therefore necessary to abandon the idea of an exact fit and to consider criteria for providing an approximate but analytical fit with a relatively low-order model.

It is proposed that a general spectrum with N maxima and N-1 included minima can be adequately represented by a pole-zero model with Z-domain transfer function

$$H(Z) = \frac{p_o + p_1 z^{-1} + \ldots + p_n z^{-n}}{q_o + q_1 z^{-1} + \ldots + q_m z^{-m}} \quad ; \quad \begin{matrix} n = 2(N-1) \\ m = 2(N) \end{matrix} \qquad (1)$$

where one complex pole pair is used to model each maximum and one complex pair of zeros for each minimum pair. This arrangement caters for the worst possible case in which highly resonant peaks are separated by transmission zeros. The possibility of there being significant minima (excluded minima) at the lower or upper frequency ends of the spectrum is discounted, since spectral information in the region of $\omega = o$ and $\omega = \omega_s/2$ (ω_s = sampled rate) is generally unreliable.

Having evolved a rule to determine the order of H(Z), the next step is to consider the criteria to be used to enforce a fit to the original spectrum. Referring again to Fig 1, it is clear that the perceptually most significant points in the spectrum are the maxima and minima, ie, $(\omega_1, A_1) \ldots (\omega_7, A_7)$.

ANALYTIC MODELLING OF SPEECH SPECTRA

These points "deserve" a better fit than any of the others. In fact, it is convenient to postulate that they deserve an exact fit, since this leads to a particularly convenient analytical solution. Thus, for the general model of equation (1), the value of H(Z) must be such that the points $(\omega_1, A_1) \ldots, (\omega_{2n-1})$ are fitted exactly and, furthermore, these should be "turning-points" (maxima or minima). Now

$$|H(Z)|^2 = H(Z)\ H(Z^{-1}) = \frac{a_o + 2a_1 \cos\omega T + \ldots 2a_n \cos n\omega T}{1 + 2b_1 \cos\omega T + \ldots 2b_m \cos m\omega T} = \frac{N(\omega)}{D(\omega)} \qquad (2)$$

where T = sampled rate.

The process of fitting $|H(Z)|^2$ to the original spectrum involves finding the "a" and "b" coefficients in the above expression so that $N(\omega)/D(\omega)$ satisfies the given criteria. The first condition implies that

$$N(\omega_k)/D(\omega_k) = A_k^2, \quad \begin{matrix} \omega_k = \omega_1, \omega_2, \ldots \omega_{2N-1} \\ A_k = A_1, A_2, \ldots A_{2N-1} \end{matrix}$$

The second condition requires that

$$\frac{d}{d\omega}\left|\frac{N(\omega)}{D(\omega)}\right| = 0 \quad \begin{matrix} \omega_k = \omega_1, \omega_2, \ldots \omega_{2N-1} \\ A_k = A_1, A_2, \ldots A_{2N-1} \end{matrix}$$

The response defined by equation (2) contains a total of 4N-1 unknowns, while the above relationships only supply a total of 4N-2 equations. In order to obtain a unique solution for the unknowns the a_o coefficient in (2) is heuristically determined. Its value may be chosen to ensure that the derived value for $|H(Z)|^2$ remains positive over its entire frequency range. The "free" parameter a_o, can in fact be used to tune the spectral fit. It is convenient to make a_o proportional to the average energy of the spectrum. With a suitable value for a_o, the other 4N-2 unknowns may be found by matrix inversion or some other analytical method.

Having found the "a" and "b" coefficients in the expression for $|H(Z)|^2$, the problem is to determine the model transfer function from this. Since

$$|H(Z)|^2 = H(Z)\ H(Z^{-1})$$

the roots of the numerator and denominator polynomials of $|H(Z)|^2$ will occur in reciprocal form. Thus every root inside the unit circle will have a corresponding root outside. A stable configuration for H(Z) may be obtained by selecting those roots of the combined function which lie inside the unit circle. Hence H(Z) may be obtained as

$$H(Z) = K\ \frac{(Z-Z_1)(Z-Z_2) \ldots (Z-Z_n)}{(Z-P_1)(Z-P_2) \ldots (Z-P_m)}$$

where P_n and Z_n are the model poles and zeros, and K is a scaling factor which

is evaluated to restore the amplitude information "lost" in finding the roots of the two polynomials.

Fig 2 shows how the derived model fits the original spectrum of Fig 1. Fig 3 illustrates the effect of the value of the a_0 coefficient on the sharpness of the model response. The consequences of an incorrect choice for a_0 is that one or other, or both, of the polynomials in $|H(Z)|^2$ may have roots on the unit-circle in the Z-plane which means that $|H(Z)|^2$ may change sign by passing through either zero or infinity. In this situation it is not possible to obtain a stable configuration for H(Z). Such a case is illustrated in Fig 4. It may be noted, however, that the matching criteria, as stated, are still fulfilled. Whenever this situation arises (in about 8% of spectra derived every 10 ms) a new attempt is made to fit the input spectrum by varying a_0 slightly about the nominal value. This increases the success rate of the fitting process to just over 99%.

3. SPEECH RESYNTHESIS For the resynthesis of utterances, the time-varying response of the speech model is obtained by finding a sequence of H(Z) which fit the input spectra, derived every 10 ms. These spectra are derived by cepstral-smoothing of the output of a bank of 128 fourth-order, narrow-band, bandpass filters. Any spectral peak with a relative amplitude of less than -50 dB is omitted from the matching process. The time movement of the poles and zeros of the derived sequence of H(Z) are then subjected to low-pass filtering (smoothing).

For the actual speech resynthesis, two options are available. H(Z) may be expressed as a product of biquadratic factors. This is equivalent in form to the serial formant synthesis method. Alternatively H(Z) may be expressed in terms of its partial fraction expansion. This method is equivalent in form to parallel formant synthesis. Both these realisations were implemented on a programmable digital speech synthesiser (PDSS) (4).

With either synthesis configuration, the filter structure is excited by an impulse train whose pitch-period reflects that of the original speech. The value of pitch, together with the digital filter coefficients, are updated every 10 ms.

4. DISCUSSION From informal listening tests, the synthetic speech produced by this system has been judged to be of fairly good quality. It does, however, possess a slight roughness. It is felt that further improvement could be attained by a pitch-synchronous update of filter coefficients. This is not possible with the present PDSS. It is interesting to note that, on the basis of a single utterance, the parallel filter configuration provides slightly better quality of speech. At present, there is no apparent reason for this. It should be pointed out, however, that, while the steady-state properties are the same for both, the dynamic properties of each are clearly different. Finally, before the resynthesis strategy can be applied to all classes of speech, it will be necessary to use techniques for voiced/unvoiced classification, and for determining the degree of voicing for mixed excitation. Preliminary analysis suggest that the technique can also be applied successfully to these types of speech.

ANALYTIC MODELLING OF SPEECH SPECTRA

REFERENCES:

1. B.S. ATAL and S.L. HANAUER 1977 J. Acoust. Soc. Amer., Vol. 50 637-655. Speech Analysis and Synthesis by Linear Prediction of the Speech Wave.

2. P.M. SEEVIOUR, J.N. HOLMES and M.W. JUDD April 1976 Proc. 1976 IEEE Int. Conf. on Acoustics, Speech and Signal Proc., 76CH1067-8 ASSP 690-693. Automatic Generation of Control Signals for a Parallel-Formant Speech Synthesiser.

3. J.P. OLIVE 1971 J. Acoust. Soc. Amer., Vol. 50 661-670. Automatic Formant Tracking by a Newton-Raphson Technique.

4. R. LINGGARD and F.J. MARLOW Oct. 1979 Computers and Digital Techniques, Vol. 2 No. 5, 191-196. Programmable, Digital Speech Synthesiser.