Proceedings of The Institute of Acoustics

A SIMPLE 4-FORMANT SPEECH ENCODER

H Christopher Longuet-Higgins (1) and John C Manley (2)

(1) Centre for Research in Perception & Cognition,
University of Sussex, BN1 9QG
(2) Hewlett-Packard Research Laboratories, Filton Road,
Stoke Gifford, Bristol BS12 6QZ

Introduction

Ever since the discovery of sharp peaks in the acoustic spectra of vowel sounds, phoneticians and speech engineers have endeavoured to encode speech as a succession of clicks and hisses modulated by a small number of narrow resonances. Among the early proponents of this approach to speech coding were Walter Lawrence (1953, 1962), Gunnar Fant (1960) and John Holmes (1962, 1982), who discussed and studied the relative merits of serial and parallel formant synthesizers. There is much to be said on both sides of that debate but the major problem confronting formant coding has always been the automatic extraction of formants from natural speech in the first place (Anthony & Lawrence 1962).

At the same time as efforts were being made to improve formant trackers and synthesizers, a rather different kind of coding system made its appearance, namely the linear predictive coding (LPC) scheme of Atal and Hanauer (1971). Linear predictive coding has the great merit that the predictor coefficients can be extracted quite automatically from reasonably noise-free speech, but their physical interpretation in terms of the shape of the vocal tract and its principal resonances must be taken with a pinch of salt - particularly as regards the frequencies and bandwidths of the higher formants. For vowel sounds the computation of the two lowest frequency formants is fairly stable, but for other speech sounds the computed formant frequencies and bandwidths jump about in an erratic manner between one 10-msec frame and the next. This fact suggests that the higher formants supplied by, for example, a 12-pole LPC analysis are not to be trusted either as objective descriptors of the acoustics of the vocal tract or as perceptually significant quantities in their own right. If this is indeed the case, then the parameters which one supplies to a formant synthesizer, whether serial or parallel, should be regarded (with the honourable exception of a vowel's f1 and f2) as quantities whose function is essentially to mimic other perceptually salient features of the short-term spectrum or autocorrelation function. Enough is known, in any case, about the perception of speech sounds, to account for the relative indifference of the ear to the precise frequencies and bandwidths of high frequency formants (in so far as these can be endowed with any objectivity) as compared with those of frequency less than about 2khz.

With these thoughts in mind, and after numerous false starts, we have developed a 4-formant speech encoder which "takes seriously" the two lowest-frequency resonances supplied by a 12-pole LPC computation and supplements them with two more, designed to reproduce faithfully the first few samples of what one may call the "residual" autocorrelation function. We explain the details of the method in the next section; all one need say at this point is

that it amounts, essentially, to a computation, for each 10-msec frame, of the frequencies and bandwidths of just 4 "formants", together with the usual parameters of gain, voicing and pitch. In these respects it resembles a straightforward 8-pole LPC encoding system, but with the important difference that the two lowest formants are more accurately identified and that the speech produced by resynthesis from the relevant parameters sounds distinctly clearer - to judge from the few utterances we have had time to resynthesize.

2. Method

The method leans heavily on the theory of linear predictive coding (as described for example in Markel and Gray's excellent monograph Linear Prediction of Speech 1976). It is similar in spirit to that proposed by H. W. Strube (1980), but involves a lot less computation. The application of the method to speech resynthesis also requires algorithms for distinguishing speech from silence and voiced from unvoiced speech, and for determining the pitch period in voiced sections.

The speech is low passed below 5khz, sampled at 10khz, pre-emphasized at $z = 0.9$ and windowed every 10 msec with a half-sine-wave window of length 200 samples. The first 13 samples of the autocorrelation function for each window are used for determining 12 linear predictor coefficients from which are obtained 12 poles in the complex plane. Some of these poles may lie on the real axis, but experience shows that during speech sounds at least 4 of the poles belong to complex conjugate pairs. The two lowest frequency resonances (pole pairs) may be identified as the first and second formant in clearly articulated vowels, but we shall in any case take the liberty of referring to them as f1 and f2.

Figure 1 indicates schematically the layout of the system. The serial order of the filters labelled f1, f2, f3 and f4 is unimportant, but for convenience we may think of f3 and f4 as preceding f1 and f2. The autocorrelation function ACF12 of the emergent speech may be used for computing the set of predictor coefficients LPC12, and from these we obtain the frequencies and bandwidths of f1 and f2. Factorising the corresponding terms out of the polynomial LPC12 we obtain an eighth-order polynomial LPC8, from which in turn is derived the autocorrelation function ACF8 of the hypothetical signal emerging from the filters f3 and f4. The symbol ACF8 indicates that the first nine samples of this function are available, but there is no need for them all in the subsequent computation, which only requires the values of the first five. We use these to calculate the coefficients of a fourth-order polynomial LPC4, from which the frequencies and bandwidths of f3 and f4 can then be obtained, if required. But if it is only required to calculate the eight filter coefficients of the entire synthesizer, this can be achieved by multiplying the polynomial LPC4 by two second-order polynomials corresponding to f1 and f2. The resulting eighth-order polynomial may be referred to as LPC8', to distinguish it from the one obtained by dividing LPC12 by f1 and f2. Its eight coefficients are those required for the resynthesis of an encoded utterance.

A SIMPLE 4-FORMANT SPEECH ENCODER

3. Results and discussion

The method was applied to seven utterances of six-figure English numerals such as one hundred and twenty three thousand, four hundred and fifty six, recorded under studio conditions, low-passed below 5khz and sampled at 10khz. ACF12 was determined at intervals of 10 msec, and the coefficients LPC12 and the first two formants f1 and f2 were computed by standard methods. f3 and f4 were also calculated, for interest, though their values are not needed for the resynthesis.

The resyntheses were carried out using for each pitch period a multi-pulse excitation derived from one of the original utterances. Comparisons were made between (i) the original utterances, as low-passed and digitised at 10khz, (ii) the utterances as resynthesised (with subsequent de-emphasis) using all 12 of the coefficients LPC12, (iii) conventional LPC8 resyntheses, derived from the ACF12 values and (iv) the utterances as resynthesised (and subsequently de-emphasised) by the method described here. Of these four sets of utterances (ii) are nearly as clear and realistic as (i); (iv) are not quite as realistic as (ii), but are distinctly clearer than (iii), which require the same number of parameters in their representation. The superior clarity of (iv) over (iii) is offset to a slight extent by faint "noises off" arising from kinks in the formant tracks, but these can be suppressed to some extent by smoothing operations (Rabiner, Sambur & Schmidt 1975) on the individual tracks.

Remarkably enough, in all the voiced frames of all the utterances, all 4 formants were represented by complex conjugate pole pairs. It is possible, in fact, to obtain entirely intelligible resyntheses using only one formant in addition to f1 and f2. But the 3-formant resyntheses so obtained have slightly odd-sounding fricatives, as one might expect from the paucity of information used for representing the high-frequency part of the spectrum.

It is perhaps worth stressing that the computation of the 4 formants is entirely uniform and automatic, like that of the LPC12 coefficients from which they are derived. Whether or not 4-formant encoders of this type eventually prove advantageous for speech transmission and synthesis by rule, one might hazard the view that the quite good quality attainable by this method supports the idea that what the ear cares about in the upper regions of the spectrum is not so much the precise frequencies or bandwidths of putative formants but the first few samples - or should one say the short-term part - of the residual autocorrelation function. Neurophysiological mechanisms for the extraction of such information from the speech wave are not difficult to conceive of, and so it may be of interest to have demonstrated that a fairly simple-minded encoder which preserves this information intact performs its task so satisfactorily.

A SIMPLE 4-FORMANT SPEECH ENCODER

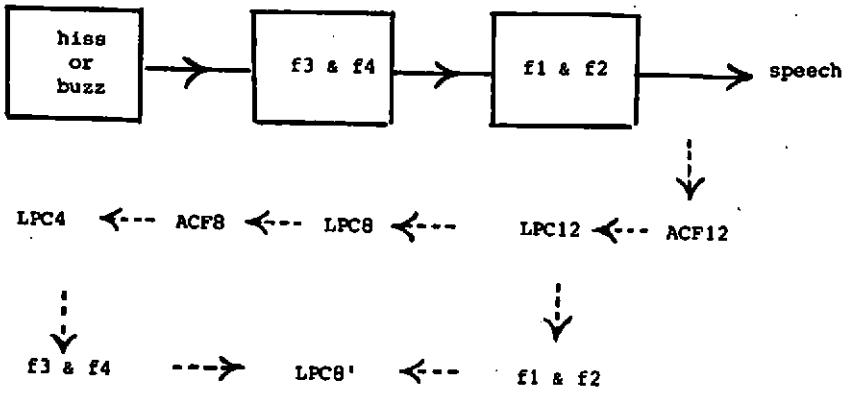

Fig. 1: the 4-formant encoder.

REFERENCES

[1] Anthony, J., & Lawrence, W. (1962). "A Resonance Analogue speech Synthesiser." Proc. Fourth International Congress on Acoustics, Copenhagen.

[2] Atal, B.S., & Hanauer, Suzanne L. (1971). "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave." J. Acoust. Soc. Amer. 50. 637-655.

[3] Atal, Bishnu S., & Rabiner, Lawrence R. (1976). "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition." IEEE Transactions on Acoustics, Speech and Signal Processing. ASSP-24. 201-212.

[4] Fant, G. (1960). The Acoustic Theory of Speech Production. s'Gravenhage, Mouton & Co. The Hague.

[5] Hess, W. (1983). Pitch Determination of Speech Signals. Springer-Verlag, Berlin.

[6] Holmes, J.N. (1982). "Formant Synthesizers: Cascade or Parallel?" JSRU Research Report No. 1017. Joint Speech Research Unit, Cheltenham, U.K.

[7] Holmes, J.N. (1962), "An Investigation of the Volume Velocity Waveform at the Larynx During Speech by Means of an Inverse Filter", Proc. Fourth International Congress on Acoustics, Copenhagen.

[8] Lawrence, W. (1953). "The Synthesis of Speech from Signals which have a Low Information Rate." Proceedings of the 1952 Symposium on the Applications of Communication Theory (ed. W. Jackson), 460-469. Butterworth Scientific Publications, London.

[9] Lawrence, W. (1962)"Formant Tracking by Self Adjusting Inverse Filtering." Paper presented at the Speech Communication Seminar, Stockholm, 1962.

[10] Markel, J.D. & Gray Jr., A.H. (1986), Linear Prediction of Speech, Springer-Verlag, Berlin.

[11] Rabiner, L.R., Sambur, M.R., Schmidt, C.E. (1975). "Applications of a non-linear smoothing algorithm to speech processing." IEEE Transactions on Acoustics, Speech & Signal Processing. ASSP-23, 552-557.

[12] Strube, H.W., (1980). "Linear prediction on a warped frequency scale." J. Acoust. Soc. Amer., 68, 1071-1076.