The Formator: a speech analysis-synthesis system based on formant extraction from linear prediction coefficients

L.L.M. Vogten and L.F. Willems

Introduction

It is well known in speech that the value of the waveform at a given instant is closely correlated with its values at previous instants, and hence represents redundant information (Flanagan, 1972). Among the many models describing the speech signal more efficiently the production model based on the linear predictability of the speech wave has been quite successful (e.g. Fant, 1960; Atal and Hanauer, 1971; Sambur, 1975). This Linear Predictive Coding (LPC) of speech represents the wave form in terms of relatively slowly varying parameters which are related to the transfer function of the vocal tract and to the characteristics of the speech source. The LPC analysis is in fact the calculation of an Mth-order digital filter, the coefficients of which are determined by minimising the mean squared error between the actual input sample and an Mth-order linear prediction of the input sample. From these M coefficients the speech wave can be resynthesised as the output of the inverse filter with the same M coefficients, excited by pulses or by noise (Markel and Gray, 1976).

The LPC method has the advantage that only relatively short segments of the speech wave are analysed in the time domain. No Fourier transform is performed and analysis can be rather fast. Unfortunately the M filter coefficients are less suitable for further processing because small errors in the coefficients can result in large errors or even instability of the inverse filter used for the synthesis.

On the other hand, we know that a description of the speech wave in terms of natural frequencies of the vocal tract or formants, is a very efficient one. Formants also change relatively slowly with time (Flanagan, 1970). Hence, if we are able to determine the formants from the M filter coefficients the LPC analysis cuts both ways.

The present contribution describes such an analysis-synthesis system based on formant extraction from the linear prediction coefficients. The system determines 5 formants from a 10th-order LPC analysis. This "Formator" has been developed at our institute and provides a powerful tool in phonetic research, because formant- (and also pitch) trajectories can be isolated, varied, stylised or quantised and the effect of these manipulations on the perception of speech can be studied. The "Formator" may also prove useful in application fields such as voice response units, low bit rate vocoders, speech recognition, etc.

The analysis part of the "Formator" has been implemented in software on our P9202 computer. The synthesis part is a digital hardware synthesiser. First we give a general description of the analysis part, followed by details of the LPC analysis, the formant extraction and the pitch extraction, whereupon the hardware synthesis

part is briefly described. The second part of this contribution gives some examples of the practical use of the system for bit rate reduction and for its use as "Intonator" (Willems, 1966) in phonetic research.

The 'Formator'

A block diagram of the system is shown in Fig. 1. The original speech is digitised with 10 kHz sample frequency, 8 bits per sample and stored on disc with the Speech Editing System (Willems and de Jong, 1974). Then an LPC analysis program is run, yielding the coefficients of a 10th-order digital filter and the amplitude parameter. From these 10 coefficients 5 second-order filter coefficients are calculated (each with 2 coefficients). The pitch period and the voiced/unvoiced parameter are determined in a separate program. These 13 parameters are then fed to the digital hardware synthesiser (Rockland 4512) and the remade speech is available. Further details of the system are described in the following sections.

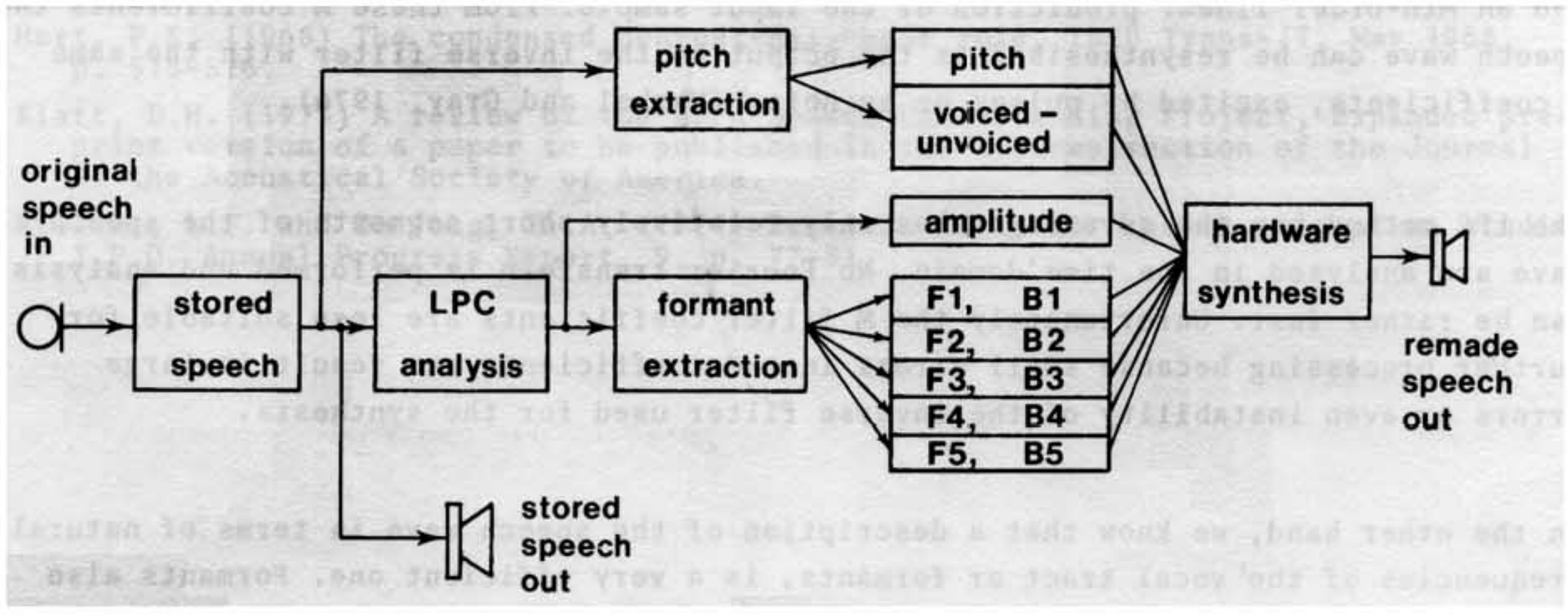

Fig. 1. Block diagram of the "Formator".

The LPC analysis

From the stored speech a 25 msec segment (250 samples) is triangularly windowed and pre-emphasised by a first-order filter $1-\mu z^{-1}$ with $\mu=0.90$. Then the 10 filter coefficients are determined by solving a set of 10 simultaneous equations which results from a least squared criterion for the error between the actual and the predicted input sample of the speech wave. In fact the autocorrelation method is used (Makhoul, 1975; Markel and Gray, 1976). After some further calculations which will be described in the next section, the analysis window is shifted over 100 samples to the next 25 msec speech segment. Thus, every 10 msec the amplitude and the 10 filter coefficients are updated. This frame rate of 100 Hz is a suitable value for normal speech; in many cases steps of 20 or 30 msec also give good results.

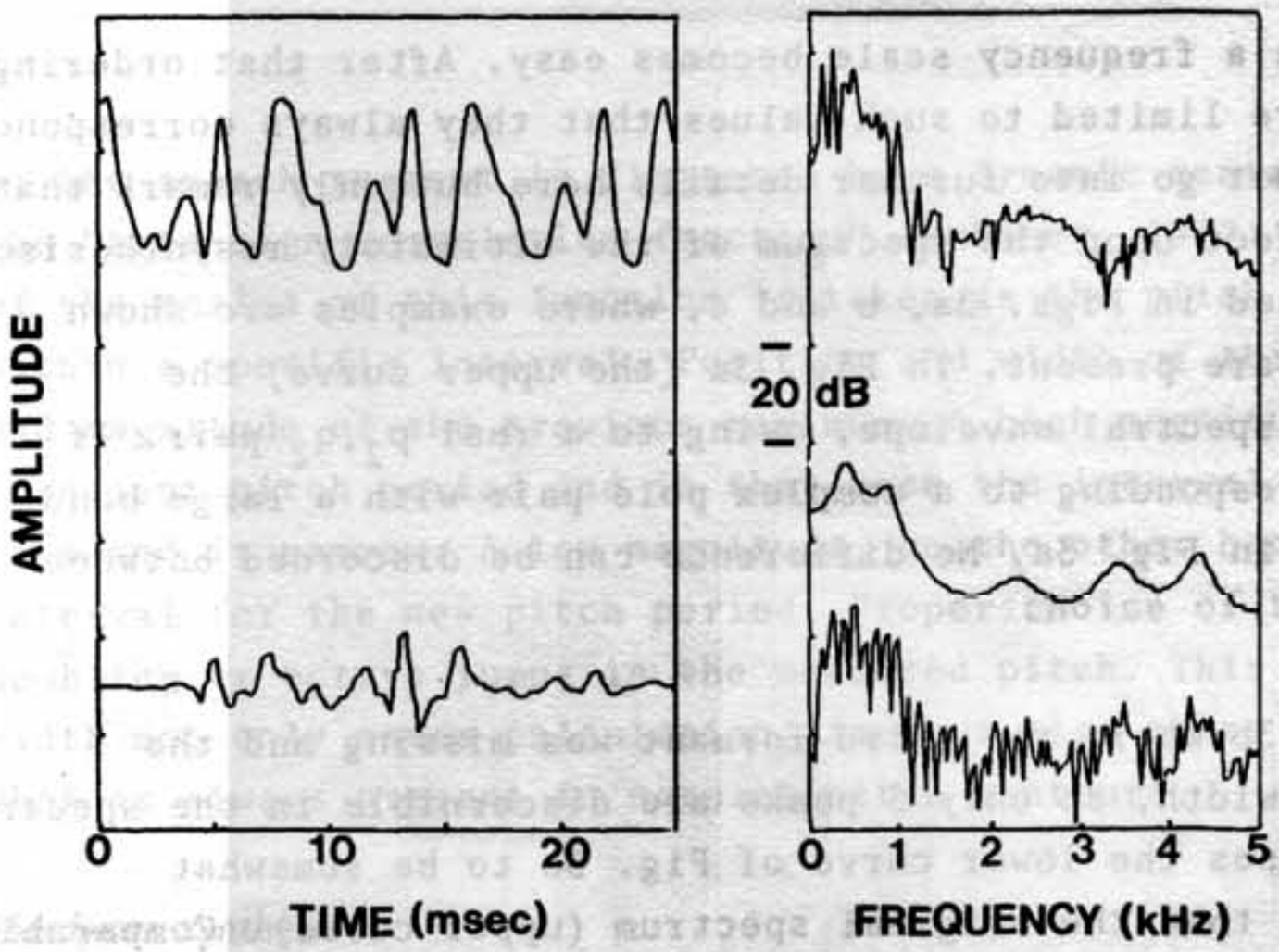

Fig. 2. Time (left panel) and frequency (right panel) representation of a 25 msec speech segment (one frame). The upper curve in each panel concerns the unwindowed signal, the lower curve is the triangularly windowed and pre-emphasised signal. In the middle of the right panel is shown the spectrum that results from the LPC analysis.

An example of the analysis result of one frame is shown in Fig. 2. The 25 msec speech segment, shown at the top of the left panel is windowed, pre-emphasised and then plotted at the bottom of the left panel. For display purposes the corresponding fast Fourier transforms of the two time signals are plotted in the right panel. This FFT is not used in the LPC calculations. The 10 predictor coefficients, which in fact represent the impulse response of the digital filter, are calculated and the corresponding FFT spectrum (also calculated for display purposes only) is shown in the middle of the right panel. It illustrates how the spectral envelope of the filter fits that of the (lower) speech wave. For this voiced speech segment (the English vowel /ɔ/ of the word "call") the 5 formants are easily discernible.

Formant extraction from the LPC data

The digital filter determined with the LPC analysis program is characterised by 10 filter coefficients $\{a_k\}$ and can be presented in the z-domain by

$$A(z) = 1 + \sum_{k=1}^{10} a_k z^{-k} \qquad (1)$$

The polynomial (1) can also be written as a product of 5 quadratic terms:

$$A(z) = \prod_{i=1}^{5} (1 + p_i z^{-1} + q_i z^{-2}) \qquad (2)$$

Calculation of the coefficients $\{p_i, q_i\}$ from the coefficients $\{a_k\}$ can be done numerically. Then we have a set of 5 $\{p_i, q_i\}$ combinations representing a cascade of 5 digital second-order filters equivalent to the 10th-order filter. These 5 second-order filters can now be conceived as the 5 formants that we are looking for.

However, we are still left with two problems: (a) the pairs $\{p_i, q_i\}$ resulting from the calculations are not naturally ordered on a frequency scale, while the formants definitely are and (b) it is possible that, especially for consonants or fricatives, one or more of the pairs $\{p_i, q_i\}$ represent a filter whose poles are real. In that case we can not speak of a formant having a tuning frequency and a bandwidth.

These problems are solved by the application of a transformation procedure to the

$\{p_i,q_i\}$ pairs so that ordering on a frequency scale becomes easy. After that ordering procedure p and q of each pair are limited to such values that they always correspond to complex pole pairs. We shall not go into further details here but only remark that these changes have no audible effect upon the spectrum of the ultimately resynthesised speech segment. This is illustrated in Figs. 3a, b and c, where examples are shown of spectra in which only 4 peaks are present. In Fig. 3a (the upper curve) the second formant is missing in the spectral envelope, owing to a real p_2,q_2 pair. If we force this pair to values corresponding to a complex pole pair with a large bandwidth the lower spectrum results in Fig. 3a. No difference can be discerned between the upper and the lower curves.

Another example is shown in Fig. 3b where the third formant was missing and the higher formants have a large bandwidth, so only 3 peaks are discernible in the spectrum. Making the p_3,q_3 pair complex causes the lower curve of Fig. 3b to be somewhat steeper in the region above 4 kHz than the original spectrum (upper curve). Comparable results are shown in Fig. 3c, where the fifth formant was missing.

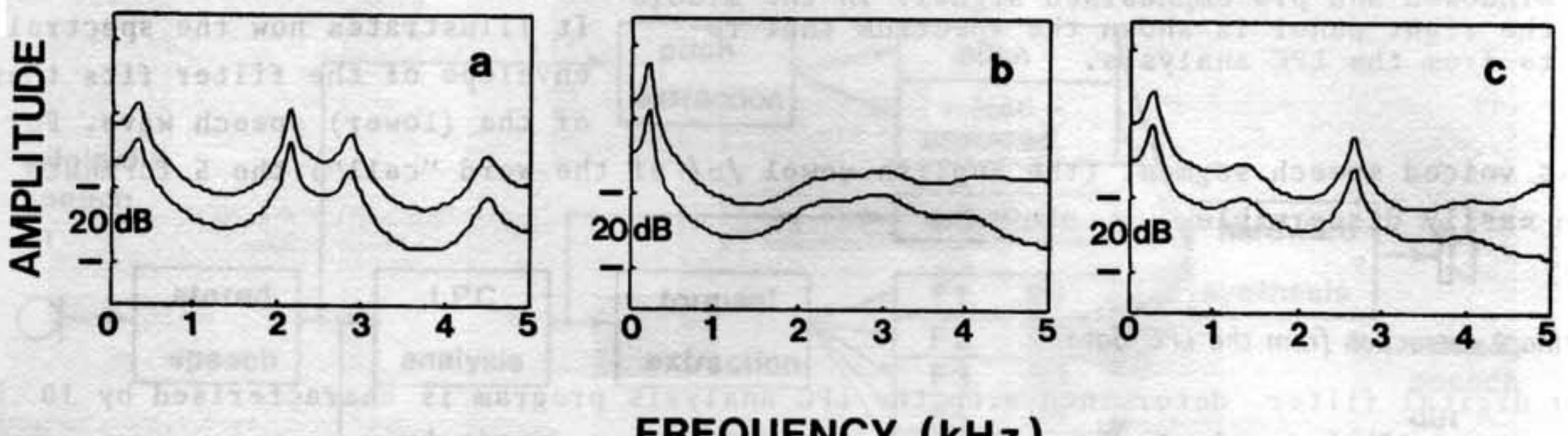

Fig. 3. The spectrum of the digital filter resulting from the LPC analysis before (upper curves) and after (lower curves) the formant extraction. In panel a the second formant is "missing", in panel b the third and in panel c the fifth formant (upper curves). They are "artificially" added by changing the second-order filter coefficients so that the real poles become complex, yielding a formant with a large bandwidth (lower curves).

These examples illustrate that the errors introduced by the "forced complex making" procedure have little or no effect upon the spectral envelope of the resulting digital filter. The result is now that we always have 5 and only 5 formants and after the definite assignment of numbers 1 to 5 inclusive, the complex and ordered $\{p_i,q_i\}$ pairs are used to determine the input parameters for the digital hardware device in order to synthesise the speech wave.

Pitch extraction

The pitch period or fundamental frequency and the voiced/unvoiced decision are determined every 10 msec from a speech segment of 35 msec. This segment is long enough to ensure that at least 2 pitch periods are present in the waveform. For the pitch extraction we use a modified version of Sondhi's (1968) method. First the spectrum

of the speech segment is flattened by a dynamic centre-clipping procedure and then the "auto-sign-correlation function" (Rabiner, 1977: method 6) is calculated. One of the maxima of this function is taken as the pitch period, provided it is positioned within a specific interval. Position and width of this interval depend on position and magnitude of the previous maximum. A high magnitude of the previous peak implies a salient pitch period and in that case the interval within which the new peak has to be found is narrow. A low magnitude, on the other hand, goes with a large possible interval for the new pitch period. Proper choice of the boundaries can avoid pitch doubling or octave jumps in the measured pitch. This method of variable window width not only saves calculation time but also takes into account a certain continuity that is always present in natural pitch contours.

Hardware synthesis

The speech wave can now be resynthesised by a digital hardware synthesiser. This device consists of a cascade of 5 second-order digital filters excited by a quasi-periodic pulse (voiced sound) or by noise (unvoiced speech). It needs the amplitude, pitch period, voiced/unvoiced and formant parameters at every pitch period. Since the parameters in the analysis are calculated at 10 msec intervals, an interpolation is necessary corresponding to the actual pitch periods. These interpolated parameter values are then used as the input parameters for the synthesiser (Rockland 4512).

Practical use of the 'Formator'

This section contains a brief description of some possibilities and results of informal experiments with the "Formator". As details of the system are still being improved "objective" test results are not yet presented.

Up to now a direct comparison between the stored input speech and the resynthesised version has been performed for about 10 different speakers (male and female). Several seconds of normal speech (Dutch and English mainly, different sentences for different speakers) were analysed. The raw analysis results were smoothed with both a running median smoothing over 5 frames and a linear filtering over 3 frames (Rabiner et al.,1975). Examples of the raw and the smoothed data are shown in Figs. 4a and 4b for the sentence "I don't think it's necessary to call in the doctor". Amplitude, pitch, voiced/unvoiced parameter and formant frequencies and bandwidths are shown for a segment of 2 sec, almost the complete sentence. In the experiments the input speech was immediately followed by the two resynthesised versions, the raw and the smoothed data. Although, of course, slight differences were audible between the original and the raw or smoothed resynthesised speech, the quality of the remade speech was good.

Bit rate reduction with the 'Formator'

Once we have a description of the spectral envelope of the speech wave in terms of ordered $\{p_i, q_i\}$ pairs related to formants, it is easy to quantise these parameters and hence reduce the bit rate. In our case the digitised input speech

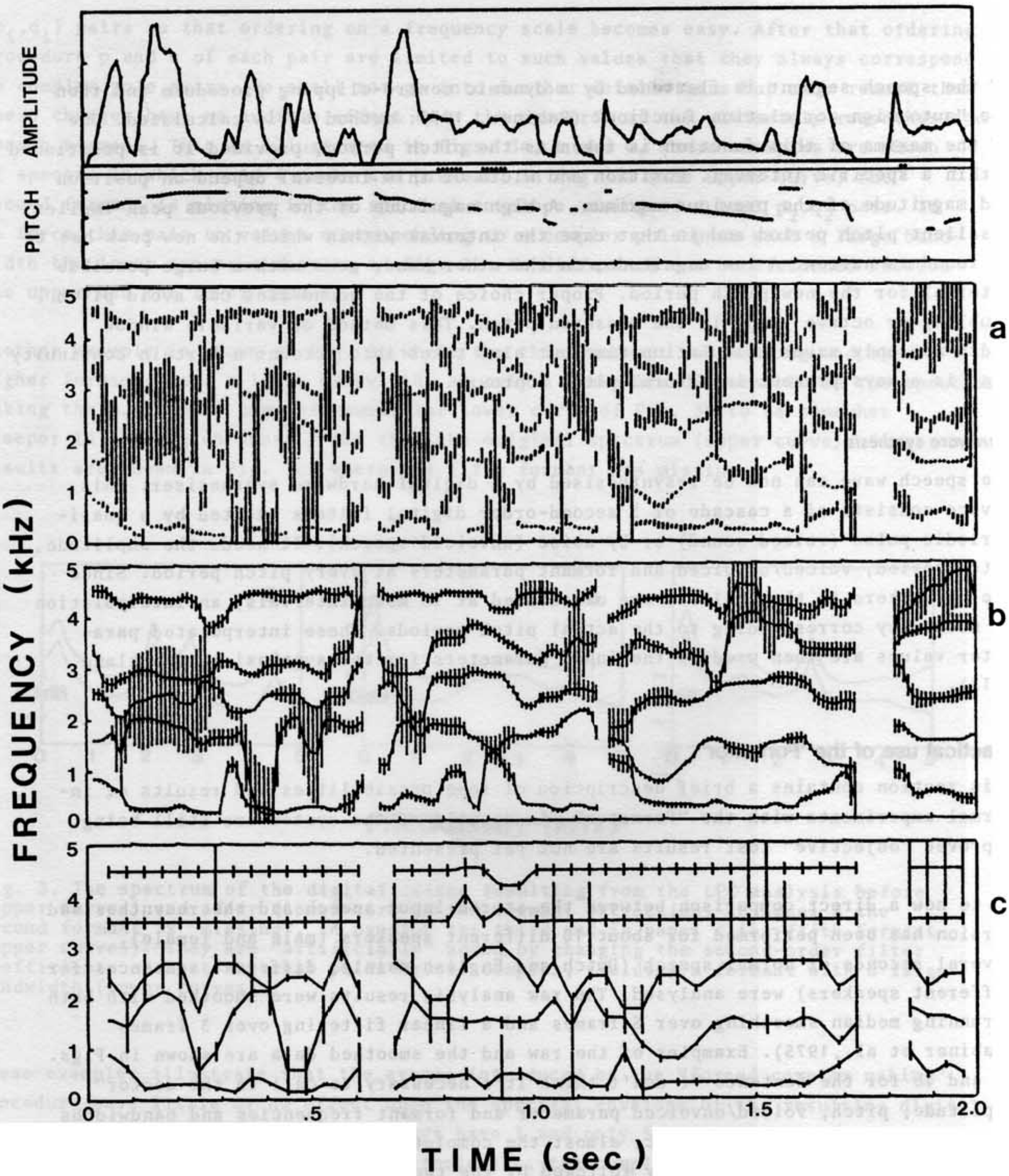

Fig. 4. The 13 parameters calculated in the analysis part of the system plotted as a function of time. Upper panel: amplitude, unvoiced marks and pitch contour. Panel (a): raw data from the analysis; the length of the vertical bars is the formant bandwidth (in Hz) divided by 2 so as not to overload the figure. The formant frequency is in the middle of each bar. Short bars indicate a narrow bandwidth and thus a sharp and high peak in the spectrum. Panel (b): smoothed formant data. Now the formant frequencies of the 10 msec frames are interconnected in order to show the formant tracks. Panel (c) the same sentence analysed at 40 msec steps and then quantised with 28 bits per frame, resulting in a bit-rate of 700 bits/sec and still acceptable in quality.

needs 80 kbits/sec (10 kHz sample frequency, 8 bits per sample). The analysed speech can be described with about 14 kbits/sec: pitch 8, amplitude 8, voiced/unvoiced 1, each formant frequency 12 and formant bandwidth 12, making 137 bits per (10 msec) frame. Preliminary experiments with quantisation of the parameters down to 28 bits per frame (amplitude 3, pitch 6, voiced/unvoiced 1, F1 up to F5 with respectively 3, 4, 3, 0, 0 bits and B1 up to B5 with respectively 2, 2, 2, 1, and 1 bits) turned out to have almost no audible effect upon the resynthesised speech compared with the unquantised version. Now we have a bit rate of 2800 b/sec. Still further reduction of information content with little loss of quality can be achieved by stylisation of the formant trajectories with an approximation by straight lines. Another possibility is to increase the analysis step width from 10 msec to 30 or 40 msec. This not only considerably reduces the frame rate and hence the bit rate but also the calculation time. In Fig. 4c an example is shown of the same sentence as in Figs. 4a and b, but now analysed with frame steps of 40 msec and then quantised with the same number of bits as mentioned above. This resulted in a description of the speech with 700 bits/sec and still acceptable in quality.

The 'Formator' as 'Intonator'

Another interesting feature of the analysis-synthesis system is the possibility of stylising the pitch contours. The natural pitch contour, measured in the analysis part of the system, can easily be replaced by a stylised intonation contour of arbitrary shape. Thus the experimenter immediately obtains an impression as to which pitch movements are relevant to the overall intonation pattern (Collier and 't Hart, 1975; 't Hart and Cohen, 1973) and which are not, simply by comparing the speech with the measured pitch contour with an artificial stylised version and then judging whether they are perceptually equivalent or not.

An example of two perceptually equivalent intonation patterns is given by 't Hart (1977). One advantage of the present system compared with previous "Intonators" (Willems, 1966) is the better quality of the remade speech.

Summary

We presented the "Formator", a speech analysis-synthesis system based on a Linear Prediction Coding of the speech wave followed by a formant extraction procedure. At present the analysis is still performed in software and a 5-formant analysis with a frame rate of 100 Hz takes about 30 times real time. Pitch is measured in a separate program, taking about 20 times real time. This "Formator" looks like becoming a promising system, not only for phonetic research but also in the field of low-bit-rate vocoders, voice response units and speech recognition.

References

Atal, B.S. and Hanauer, S.L. (1971) Speech analysis and synthesis by linear prediction of the speech wave, J. Acoust. Soc. Am., 50, p. 637-655.

Collier, R. and 't Hart, J. (1971) A grammar of pitch movements in Dutch intonation, I.P.O. Annual Progress Report, 6, p. 17-21.

Fant, G.C.M. (1960) Acoustic theory of speech production, Mouton & Co, 's-Gravenhage, The Netherlands.

Flanagan, J.L. (1970) Synthetic voices for computers, I.E.E.E. Spectrum, October 1970.

Flanagan, J.L. (1972) Speech analysis, synthesis and perception, Springer, Berlin.

Makhoul, J. (1975) Linear prediction: a tutorial review, Proc. I.E.E.E., 63, p. 561-580.

Markel, J.D. and Gray, A.H. (1976) Linear prediction of speech, Springer, Berlin.

't Hart, J. and Cohen, A. (1973) Intonation by rule: a perceptual quest, J. Phonetics, 1, p. 309-327.

't Hart, J. (1977) Pitch contour stylisation on a high-quality analysis-resynthesis system, this issue.

Rabiner, L.R. (1977) On the use of autocorrelation analysis for pitch detection, I.E.E.E. Trans. ASSP-25, p. 24-33.

Rabiner, L.R., Sambur, M.R. and Schmidt, C.E. (1975) Applications of a nonlinear smoothing algorithm to speech processing, I.E.E.E. Trans. ASSP-23, p. 552-557.

Sambur, M.R. (1975) An efficient linear prediction vocoder, Bell. Syst. Techn. Journ., 54, p. 1693-1723.

Sondhi, M.M. (1968) New methods of pitch extraction, I.E.E.E. Trans. AU-16, p. 262-266.

Willems, L.F. (1966) The Intonator, I.P.O. Annual Progress Report, 1, p. 123-125.

Willems, L.F. and de Jong, Th.A. (1974) Research tools for speech perception studies, I.P.O. Annual Progress Report, 9, p. 77-81.