Proceedings of The Institute of Acoustics

AUTOMATIC FORMANT ANALYSIS

B C DUPREE

JOINT SPEECH RESEARCH UNIT, PRINCESS ELIZABETH WAY, CHELTENHAM, GLOUCESTERSHIRE. GL52 5AJ.

Introduction

It has been known for some time (1) that, when controlled by suitable hand derived parameters every 10 ms at a rate of approximately 5500 bits/s, the JSRU parallel formant synthesizer is capable of producing speech of high quality. Seeviour et al (2) have described an automatic computer simulated formant analyser which produces control signals for this synthesizer from natural speech. The quality of speech re-synthesized using such automatically derived control parameters is by no means as good as that produced by hand derived parameters. Further work has been performed on this analyser in an attempt to improve its overall performance on a range of talkers.

To obtain formant parameters, the analyser uses an analysis-by-synthesis method in which the log power spectrum of a short section of the input speech is compared with log power spectra of several versions of the synthetic speech signal. These signals are generated, within the analyser, by a simplified model of the JSRU synthesizer ultimately used to produce the re-synthesized speech. During voiced sounds the analysis is performed over a 3.2 ms interval immediately following glottal closure when the vocal tract is ideally in force free oscillation (3). The pitch synchronous effects of sub-glottal coupling are thereby avoided during analysis, unless the fundamental frequency approaches 300 Hz.

Initial Analysis

The speech signal is first band-limited to 5 kHz and sampled at 10 kHz. A fairly simple peak picking algorithm is then used to indicate the points of major excitation of the vocal tract, ie the epochs of glottal closure during voiced sounds. During unvoiced sounds, the supposed excitation points generally correspond to regions where there is a local increase in the power of the speech signal. A log power spectrum is then computed from 32 speech samples (60 samples if unvoiced) following the excitation points in each 10 ms frame of the input speech. If there is a succession of unvoiced frames then the spectrum is modified by including a proportion of the spectra of all previous unvoiced frames in the series. Overall, it appears that this method of dealing with unvoiced sounds is more robust than the previous method which involved taking the autocorrelation of a longer section. This is particularly so when errors occur in the voicing decision and during sounds with mixed excitation. The benefit presumably arises because in the present scheme some semblance of the pitch synchronous procedure is retained in such circumstances.

Next, the log power spectrum is examined in order to determine an initial allocation of formant frequencies. A spectral peak picking procedure is applied to indicate the major formant peaks. In general, there will be more spectral peaks than there are formants to be allocated to them and so up to 6

different versions of the allocation of the 5 formants (Nasal, 1st, 2nd, 3rd and 4th formants) are retained for further investigation.

Refinement of Estimates

For each of the above 6 versions the formant amplitudes must be calculated and the frequencies refined. A log power spectrum is computed for each formant and compared, in the region of the formant peak only, with the log power spectrum obtained from the input speech. This comparison provides an estimate of the formant amplitude together with a suitably weighted spectrum error measure. This procedure is performed repeatedly with the frequency of the formant generator set to a number of values near the initial allocation. That formant frequency giving the smallest measured error is chosen. This process does not take account of the interaction between adjacent formants. The complex spectra resulting from the chosen formant frequencies (and calculated amplitudes) of a version are then combined and one full log power spectrum produced for that version.

There then follows a two cycle iterative process which further refines the estimates for each version so as to take account of the interaction between formants. The spectrum due to a formant is computed, the formant frequency being set to various values in the region of the previous value. An estimate is made of the effect of inaccuracies in the previous estimate of the full synthetic log power spectrum and, by comparison with the spectrum obtained from the natural speech, a new decision is made as to that formant's frequency and amplitude.

For use in the above procedures the complex spectrum due to an individual formant is formed by taking the Fourier transform of 32 (60 if unvoiced) samples of the output of a 2nd order recursive filter representing a formant generator. The filter input is a train of pulses representing the airflow through the glottis and the samples for the transform are those immediately following a major excitation. It has been found experimentally that this modelling of the excitation is an improvement compared with the previous model which was equivalent to replacing the pulse train with a single impulse.

In order to choose one version of 5 formant frequencies and amplitudes from the 6 versions available, each of the 6 synthetic log power spectra is compared with the natural equivalent using a suitably weighted squared error measure. The weights currently used weight only against regions where the ordinate values in the natural spectrum are low, and against frequencies higher than about 3 kHz. In particular it has been found inadvisable to give less than full weight to the lower few hundred hertz of the spectrum as this tends to allow the occasional selection of a version containing errors in the estimates of the nasal and 1st formants. Such errors are particularly annoying in the perception of synthetic speech occupying the full band from 40Hz up to 4 kHz. The above spectral goodness of fit measure is then combined with a continuity measure indicating how closely the formant frequencies for that version correspond with the choice actually made for the previous frame. The continuity measure depends on the confidence with which the previous choice was made. The combined measure is used to make the final decision.

AUTOMATIC FORMANT ANALYSIS

Formant Tracks

The formant tracks can be improved by incorporating some delay and using the results for past and future frames to apply suitable non-linear smoothing to the parameters. At present this merely consists of a rule whereby any given parameter is not normally allowed to lie outside the range formed by the values of that parameter in the immediately preceding and following frames. It seems likely that more sophisticated speech-like constraints will have a beneficial effect on speech quality. They should incorporate as much delay as possible so that decisions can be made in the light of past and future events and ideally should have their effect as early as possible in the analysis cycle. Such procedures, however, require an increase in computational effort and may tend to result in synthetic speech which is more stylized and retains fewer of the idiosyncrasies of the original speaker.

Results

Recent work has mainly used 3 sentences each spoken by 5 British talkers. Because of the large amount of computation involved the modifications were usually optimised using various short sections of speech. These sections were taken from utterances which do not occur on the tape recording presented, and thus some indication is given of the performance of the analyser-synthesizer combination. A number of subtleties of the input speech are preserved and the performance is not unreasonable during fricative sounds and sounds of mixed excitation. Most of the significant remaining errors are those which occur when it is not easy to discern where the formants should lie using spectra covering the few tens of milliseconds available to the analyser. As mentioned above, longer system delay would be needed to ameliorate these faults.

Hardware

The computer simulated formant analyser described above runs in about 3000 times real time when implemented in FORTRAN on an ICL 4-70 general purpose computer. In order properly to optimise the various details of the algorithm it is necessary to process large quantities of speech. Therefore, special purpose digital hardware is at present under development (by the Plessey Co. Ltd.) which will perform the analysis in real time and allow experimental conversation via the system.

The principles of the construction of the analyser are shown in Figure 1. Overall control is provided by a conventional minicomputer. This also allows the operator to change details of the algorithm. Long-term storage of the program is provided by floppy discs. Most of the operations are performed by approximately ten bipolar bit-slice microprocessors (cycle time 150 ns), each microprocessor having a defined task such as the spectral peak picking algorithm or the calculation of the various error measures. Two operations, however, must be performed much more rapidly than is reasonable using the microprocessors, namely the operation of the formant generators and the Fourier transforms which must be performed on the synthetic waveforms. Fortunately, these are defined tasks which are not likely to require modification in the future and so they are performed using hard wired TTL digital logic.

AUTOMATIC FORMANT ANALYSIS

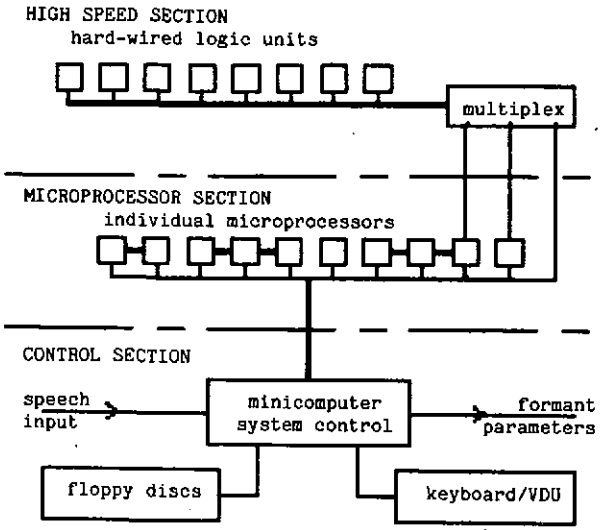

Fig. 1 Hardware analyser under development

Each of the high speed units shown in Figure 1 will perform a combined task of generating a synthetic single formant waveform and computing its Fourier transform in about 64 microseconds.

The fundamental frequency extraction will be performed by a currently available analogue device.

References

(1) J N Holmes 1973 IEEE Transactions on Audio and Electroacoustics 21, 298-305."The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer".

(2) P M Seeviour, J N Holmes and M W Judd 1976 IEEE Conference Record; Conference on Acoustics, Speech, and Signal Processing, 690-693. "Automatic generation of control signals for a parallel formant speech synthesizer".

(3) J N Holmes and E M Thornber 1973 Proceedings of British Acoustical Society Spring Meeting, paper number 73SHB5."Formant frequency measurement by waveform matching during closed-glottis periods".