A HIGH PERFORMANCE DIGITAL HARDWARE SYNTHESISER

J.E. Clark (*), C.D. Summerfield (**), R.H. Mannell (*)

(*) Speech, Hearing and Language Research Centre
Macquarie University

(**) Centre for Speech Technology Research
University of Edinburgh

ABSTRACT - This paper describes the phonetic aspects of the development of a high performance real time digital hardware synthesiser based on five formants in parallel connection, a nasal formant and a zero. The hardware realisation is based on a TMS320 digital signal processing chip, with a Z80 ancilliary processor to manage data communications and control.

INTRODUCTION

Formant synthesisers now have a long history of development. Early devices such as PAT (Lawrence (1953)) and OVE II (Fant & Martony (1962)) were constrained in parametric complexity and performance by the available technology. The idea of defining synthesis parameters by painting conductive ink on plastic sheets to control several racks full of hot and unreliable vacuum tube electronics has moved in the space of 30 years from the remarkable to the unthinkable.

Modern technology has released us from many past constraints. The objective in designing this present synthesiser has been to provide sufficient parametric flexibility in its specification of the time-varying speech spectrum to make the quality of the input parametric data the major limitation on its performance potential.

GENERAL SYNTHESISER ARCHITECTURE

The synthesiser is based on 5 formants in parallel connection, and generates speech-like signals over a frequency range from 0 - 5 KHz which is the limit set by the internal digital hardware sampling rate. It has been shown by Fant (1960) that serial connection of a set of four formant filters, together with high frequency equalisation for the absence of higher order filter poles, will provide an accurate model of vowel production acoustics in which the amplitudes of the vowel resonances are automatically defined by the frequency distribution of the resonances. Unfortunately, the benefit of such modelling fidelity is seriously offset by its inflexibility in being unable to specify formant amplitudes independently. Sounds generated with constrictions greater than those for vowels in the supraglottal tract, those with fricational excitation sources above the larynx, and those involving oral and nasal tract coupling, are not readily modelled using the serial approach. The provision of supplementary filters to the main serial formant vowel filter system to compensate for this deficiency is not an adequate solution. It is very difficult to maintain even the appearance of pole continuity in spectral structure when supplementary filters are used, with the result that perceptual integration in consonant vowel sequences may be lost. Traditional spectral distortion problems in vowel modelling with parallel connection synthesisers has been successfully overcome by Holmes (1972), so that the only real cost in choosing a parallel architecture is the additional formant amplitude parametric data it demands. The serial/parallel argument is discussed in detail by Holmes (1982). The only alternative is a hybrid solution of the kind used by Klatt (1980).

PHONETIC PERFORMANCE CRITERIA

i) Control of basic voice quality dimensions.
ii) Control of formant filters allowing the modelling of complex spectra associated with highly damped vocal tract acoustics.
iii) Provision for source/filter interaction.
iv) Inclusion of a spectral zero and additional secondary resonance.
v) Vowel resonance modelling without appreciable spectral distortion.

SYNTHESISER DESCRIPTION

The structure of the present synthesiser is shown in the schematic of Fig. 1. Its description is most conveniently considered in terms of the source and filer model it represents.

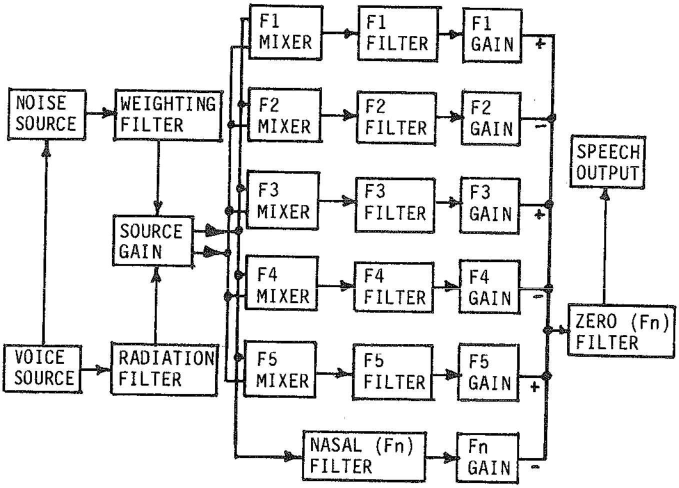

Fig. 1

a) Excitation Sources
There are two basic forms of vocal tract excitation to be modelled, periodic from larynx vibration, or aperiodic from airflow turbulence. In most synthesisers, the periodic source is an impulse train filtered to approximate an idealised spectrum of the volume velocity waveform of phonation. The present synthesiser provides an approximation to the volume velocity waveform itself, which is synthesised from polynomial function sections as suggested by Rosenberg (1971). Fundamental frequency, and the rise and fall times of the waveform (which indirectly provide control of the open to closed ratio), are all dynamically controllable. The latter pair determine the overall slope and weighting of the periodic source spectrum.

There is provision for noise insertion in the closed phase of the larynx waveform, and the introduction of dynamically variable pitch jitter, both of which produce forms of aperiodicity in the source spectrum. These periodic source parameters all contribute to the determination of overall voice quality. In addition, there is a glottal area function signal (half sinusoid) derived from the volume velocity waveform generator which can be used to modulate F1 bandwidth up to a factor of 5 times at the sine peak, and give a limited simulation of source/filter interaction.

The volume velocity waveform is further shaped by a radiation correction filter in the form of a differentiator having a variable corner frequency from 0Hz - 700Hz. The function of this filter is to give a rising high frequency characteristic simulating the effect of sound radiating from a very small area on a 9 cm diameter sphere representing the lips and head respectively. Insertion of the radiation filter at this point helps to optimise use of available dynamic range of filter calculations in the signal processing chip. Fig.2 shows the effect of varying the radiation zero frequency on a typical periodic source spectrum.

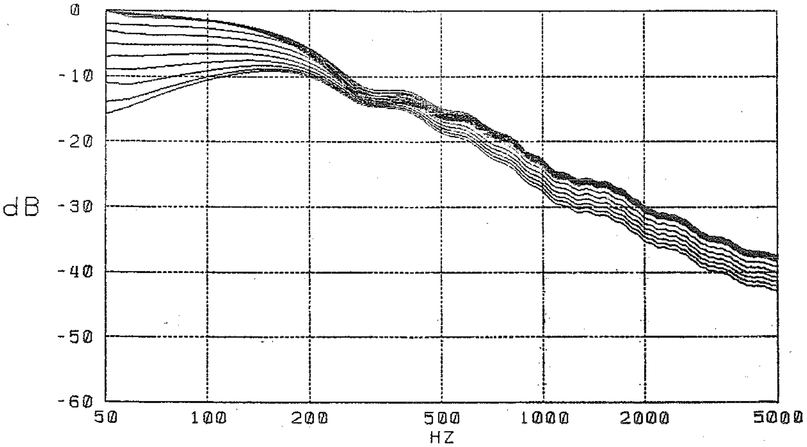

Fig. 2

The aperiodic source is itself spectrally flat, and is shaped using a weighting filter centered at 0.5 KHz. It is designed to attenuate energy below this frequency as there is little low frequency energy in fricational sounds, and provide a spectral energy slope above the centre frequency approximating that of the radiation corrected periodic source. This ensures that both sources have spectral energy distributions with compatible dynamic operating ranges for the signal processing chip. The aperiodic source may also be modulated from a depth of 0 - 100% by the volume velocity waveform to provide the source excitation required for certain forms voiced fricative.

There is an overall gain control on the excitation source outputs which sets the input amplitude to the formant filter system. At the input to each formant there are linear mixers to set the proportions of periodic to aperiodic excitation in a similar manner to that used by Holmes (1971). This allows the noise to voicing ratio to be varied across the frequency range of the formant filter system, as periodicity and aperiodicity are not always uniformly distributed across the spectrum of speech signals.

b) Filters

The synthesiser filter is based on a bank of six resonators, five to provide formant shaping for orally generated spectra, and one so-called nasal formant filter to assist in the simulation of spectra resulting from the contribution of nasal cavity resonance. All the formant filters have dynamically variable centre frequencies and bandwidths. In normal operation it is likely that only the three lowest oral resonators and the nasal resonator will be varied. Access to control of F4 & F5 is necessary where effective vocal tract length is being scaled for adult/child or male/female differences, to model particular speakers accurately, and for changes in voice quality which alter long term vocal tract settings such as larynx height. Control of F4 & F5 may also be desirable for aspects of some segmental speech structures including certain fricatives, and strongly rhotacised or palatalised sounds.

Very wide range bandwidth control from 10Hz to 2500Hz is provided in all resonators so that highly damped vocal tract conditions of the kind found in stops and fricatives can be simulated. Gain controls are provided at the output of each resonator to set formant levels, and are calibrated to give equal amplitude for equal gain settings with equal bandwidth at the standard neutral vowel centre frequencies. Output can thus be controlled via the input gain control setting, the bandwidth setting, and the output resonator gain. This allows very versatile control over rapid spectral energy changes, and potentially, more accurately modelled simulation of spectra involving large dynamic changes in vocal tract damping.

Prior to final mixing, the oral resonator outputs are passed through filters to minimise spectral distortion effects arising from interaction between the filter skirt responses. Unless this precaution is taken, changes in formant parameter settings may cause serious distortions in the combined resonator response. With appropriately chosen skirt filter characteristics, and arithmetic sign alternation of resonator outputs, the parallel connected filter will produce a combined filter response almost identical to that of their serially connected counterpart. Fig. 3 shows the spectrum of a neutral vowel generated with the periodic and aperiodic excitation sources.

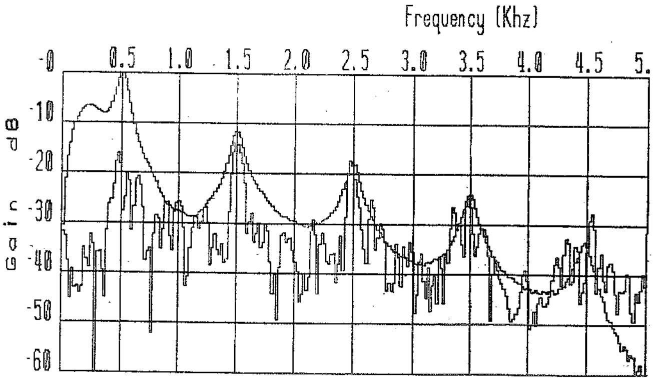

Fig. 3

Finally, the combined output of the formant resonators is passed through an antiresonance filter with a dynamically variable centre frequency and bandwidth which determines both its depth and width. This is used to produce spectral zero effects of the kind found in fricatives, approximants, and nasals.

PARAMETRIC CONTROL

There are 39 parameters which may be updated at a variable rate from 3 - 31 mSec. The start of the glottal waveform may be synchronised to the start of the parametric data frames to allow more consistent generation of rapidly changing spectral components such as stop release bursts. To avoid spurious transient effects, there is provision for interpolation between parameter updates of the gains of all formant filters, and the centre frequencies and bandwidths of F1 - F3.

PARAMETER SPECIFICATIONS

a) Voice Source
Fo - 20 to 1000Hz in 4096 increments
Glottal wave risetime - 0 to 1/Fo limit in 512 increments
Glottal wave falltime - 0 to 1/Fo limit in 512 increments
Closure phase noise gain - 0 to -72 dB in 16 increments
Fo jitter - 0 to 6% in 16 increments
F1 bandwidth modulation - 1X to 5X in 16 increments
Glottal wave start synch. with parameter frame - on/off

b) Noise Source
Voicing modulation - 0 to 100% in 16 increments

c) Source Control
Source gain - 0 to -7.2dB in 16 increments
Voice to noise source amplitude proportion at inputs of F1 to F5 - 100% to 0 in 16 increments

d) Filters
Radiation filter corner frequency - 0 - 500Hz in 16 increments
Formant filter centre frequency (F1 to F5, Fn) - 0 to 5000Hz in 512 increments
Formant filter bandwidth (F1 to F5, Fn) - 10 to 2500Hz in 256 increments
Formant filter gain (F1 to F5, Fn) - 0 to -72dB in 4096 increments
Zero filter centre frequency - 0 to 5000Hz in 512 increments
Zero filter bandwidth - 10 to 2500Hz in 256 increments

e) Parameter Controls
Synthesiser parameter frame update rate - 3 to 31 mSec in 28 increments
Gain interpolation (F1 to F5, Fn, Fz) - on/off
Centre frequency interpolation (F1 to F3) - on/off
Bandwidth interpolation (F1 to F3) - on/off

CONCLUSION

The design described has attempted to meet the 5 criteria outlined at the beginning of the paper. Certain aspects of excitation source properties and control, and of spectrum shaping, have been made sufficiently redundant to allow the exploration of alternative solutions to the simulation of complex speech spectra.

ACKNOWLEDGEMENTS

The authors would like to thank members of the S.H.L.R.C. technical staff for their assistance during this project, in particular Ray Robinson and Mark Stevens. The work was supported under Telecom Contract 68250.

REFERENCES

FANT, G. (1960), "Acoustic Theory of Speech production", (Mouton: The Hague).

FANT, G. & Martony, J. (1962), "Instrumentation for parametric synthesis" STL-QPSR, 2, 18-24.

HOLMES, J.N. (1972), "Speech Synthesis", (Mills & Boon: London).

HOLMES, J.N. (1982), "Formant synthesisers: Cascade or parallel?" JSRU Research Report No. 1017.

KLATT, D. (1980) "Software for a cascade/parallel formant synthesiser", JASA, 67, 971-995.

LAWRENCE, W. (1953), "The synthesis of speech from signals with a low information rate", in W. Jackson (ed.), "Communication Theory", (Butterworth & Co: London).

ROSENBERG, A.E. (1971), "Effect of glottal pulse shape on the quality of natural vowels", JASA, 49, 583-590.