Proceedings of the Institute of Acoustics

REAL-TIME GAMMATONE FILTERBANK SPECTROGRAPHY

David M Howard, Paul E Garner and Andrew M Tyrrell

Parallel and Signal Processing Applications Research Group, Electronics Department, University of York, Heslington, York YO1 5DD, UK.

1. ABSTRACT

Speech scientists make extensive use of the spectrograph in their laboratory work. Conventional spectrographs produce an energy/frequency/time spectrogram of the input signal, based on the principle of a fixed bandwidth filterbank analysis. Usually at least two filter bandwidth settings are available; a wide bandwidth to provide 'good' time response and a narrow bandwidth for 'good' frequency response. This paper discusses an initial implementation of a spectrograph system which models the human peripheral hearing system as a bank of GammaTone filters with bandwidths determined on an equivalent rectangular bandwidth (ERB) scale. The bandwidth of ERB-based analysis filters increases with frequency; thus the human hearing system exhibits 'good' frequency response at low frequencies and 'good' time response at high frequencies. This bank of filters operates in parallel in the model, and we have used transputers to implement a real-time waterfall display spectrograph based on ERB spaced GammaTone filters. The graphics output of the system is handled by a 3-D transputer graphics processor board developed at York.

2. INTRODUCTION

The peripheral hearing system, taken to consist of the outer, middle and inner ear to the point of neural transduction in the organ of cortii, provides an analysis of the incoming acoustic signal prior to transmission via neural links to higher centres of the brain. The critical band filter mechanism of the cochlea describes the frequency analysis carried out by the basilar membrane which gives the basis for the 'place' and 'temporal' theories of peripheral hearing (e.g. Moore, 1982). The mechanical vibration of the basilar membrane, which stimulates the hair cells of the organ of cortii to fire and generate neural impulses, relates to the output waveform from each filter. This model of the peripheral hearing system provides the basis for contemporary psychoacoustic theories of human perception of pitch, timbre and loudness.

Currently available laboratory speech spectral analysis systems are based on a fixed bandwidth acoustic analysis. The speech spectrograph (e.g. Baken, 1987) provides the user with a choice of analysis filter bandwidths, of which 300Hz (wide) and 45Hz (narrow) are most commonly used. These enable the signal to be analysed with 'good' time resolution or 'good' frequency resolution respectively. Visual displays of such patterns have furthered our understanding of the nature and role of acoustic cues in speech and hearing which provide the basis of operation for many speech synthesis and recognition systems. It is clear that current knowledge of the nature of the acoustic patterns in speech is still very lacking, for example, from the poor perceptual ratings given when the *naturalness* of synthetic speech rather than the *intelligibility* is investigated.

Acoustic analysis based on the output from a model of the peripheral hearing system would provide a graphics display of the acoustic patterns which are sent to the brain by each ear. Investigation of the nature of acoustic patterns produced by such a model has potential for

enhancing our understanding of the acoustics of speech and hearing thus enabling improvements to speech synthesis and recognition systems. Such a model could be used in display systems to provide real-time visual feedback for speech and singing training. The acoustic patterns evoked in response to other sounds such as music and environmental sounds could also be studied.

This paper describes an implementation of a model of the human peripheral hearing system which at present produces a 3-D waterfall spectral display output. The peripheral hearing system is modelled as a concurrent process and the parallel processing capabilities of the transputer enable such processes to be implemented in a more straightforward manner than would be possible with a sequential processor. The speed advantage gained by using transputers enables the performance required for real-time processing to be achieved. In this case, an array of 42 transputers is used which enables a 32 channel filter system to be implemented in real-time.

3. GAMMATONE FILTERS

The model of the human peripheral hearing system presented in this paper makes use of GammaTone filters (e.g. Patterson, 1976) to model each auditory filter. The GammaTone filter was introduced to describe the shape of the impulse response function of the auditory system as estimated by the reverse correlation function of neural firing times. Inherent in the GammaTone model is 'good' frequency analysis at low frequencies and 'good' time analysis at high frequencies which provides one of a number of its potential advantages over conventional spectrographic analysis. A bank of GammaTone filters working in parallel is the basis of the human peripheral hearing model. This inherently parallel model is implemented on a bank of transputers where each transputer, in the limit, represents a single filter.

The form of the function is that of an amplitude modulated carrier tone of frequency f_0Hz, with an envelope proportional to $t^{n-1} \exp(-2\pi bt)$, which is the Gamma distribution. The parameters of the GammaTone filter are: n, the order, which controls the relative shape of the envelope, becoming less skewed as n increases; b which controls the duration of the impulse response function, increasing b leading to shorter duration; f_0 which determines the frequency of the carrier; and ϕ the carrier phase, which determines the relative position of the fine structure of the carrier to the envelope.

The human peripheral auditory system can be modelled as a bank of overlapping GammaTone filters. As the centre frequency of the bandpass filter increases so does the bandwidth of the filter.

4. SIMULATION OF GAMMATONE FILTERS ON TRANSPUTER NETWORKS

To implement more than just a simple two channel GammaTone filter bank required a large number of transputers. A B042 board (manufactured by INMOS) was used. This contained 42 T800 transputers arranged in a 6 X 7 grid array. The transputers are connected by their links in a mesh. The transputers on the B042 board have no external memory fitted nor was it possible to add any external memory. Thus, any code executing on the transputers must fit within the internal 4 kByte memory.

REAL-TIME GAMMATONE FILTERBANK SPECTROGRAPHY

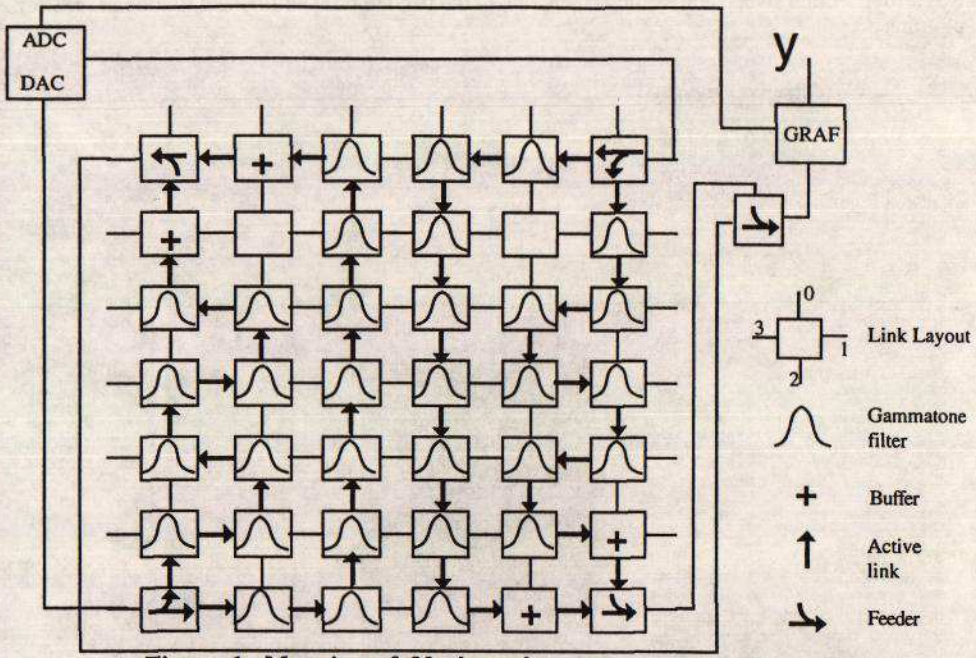

Figure 1: Mapping of 32 channel system

The mapping used for placing 32 channels onto the B042 board is shown in Figure 1. The mapping consists of four pipes each containing eight filters. Four pipes are used so that the links between processors are not flooded by the information flow along each pipe. Each filter has fed to it an array of nine integers. The first element of the array contains the most recent input sample, each of the other elements are used for the results from each filter, the array being progressively filled as it passes through the pipe. The data rate along each line is therefore 9 elements x 2 bytes x sampling rate. At a sampling rate of 12 kHz this gives a uni-directional data rate along each link of 216 kBytes/S, meaning that each transputer in the pipeline must read/write 432 kBytes/S. This is well within the 1740 kBytes/S possible using a T800 with links speed set to 20 Mbits/S. However, pipelines containing more than around twelve filters are not feasible since the transputer cannot simultaneously read/write large arrays and carry out filter calculations.

5. RESULTS

The initial version of the system (Tyrrell et al., 1992; Swan et al., 1994) enabled a real-time spectrogram as well as a real-time spectral bargraph display to be produced. A screen dump of an example spectrogram is shown in figure 2 for the vowels [i:], [a:] and [u:] spoken by an adult male. It should be noted that the frequency range of the plot is 50Hz-3kHz and it is plotted on an ERB scale. However, the display quality was crude and the best time scale which was achieved

was a four second sweep across the screen, due to the need to transfer the data to the host PC for plotting.

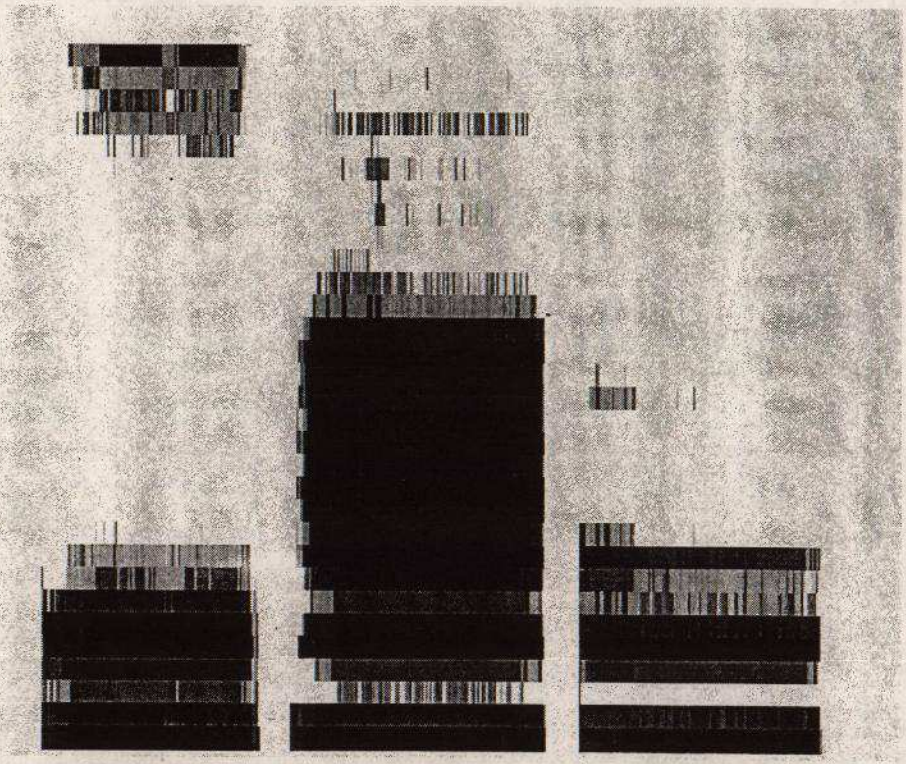

Figure 2: GammaTone spectrogram for [i:], [a:] and [u:] spoken by an adult male. (NB Frequency plotted on an ERB scale.)

To increase the usability in terms of the minimum sweep time and the crudeness of the display of this system, a 3-D transputer-based graphics system was designed and constructed. This graphics board made use of two transputers and it has facilitated interfacing to the processing network thus avoiding a data bottleneck. A block diagram of the hardware of this board is shown in figure 3. A number of standard software routines were written to enable the outputs from the processing network to produce 3-D graphics. The performance of the 3-D graphics board is summarised in table 1.

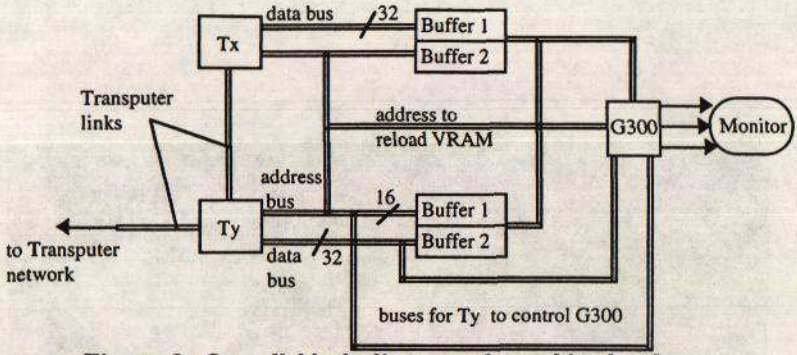

Figure 3: Overall block diagram of graphics hardware

Measurement	Lines per second	Frames per second	Points per second
Graph of test data, clearing screen before each frame, using display processors	5878	1.48	6062
Graph of test data, without clearing screen, using display processors	6214	1.56	6390
Graph of test data, not using display processor, ie no drawing or screen clearing	13922	3.5	14336

Table 1: Performance of 3-D graph of test data

A waterfall spectral display is available from the new graphics board at present and an example is shown in figure 4 for the vowels [i:], [a:] and [u:] spoken by an adult male. The frequency range of the plot is 50Hz to 3kHz and it is plotted on an ERB scale. It should be noted that this figure has been produced by photographing the graphics screen as it is not currently possible to digitally capture a screen from the transputer graphics board.

The different formant positions for the three vowels can be seen and compared with the spectrographic version in figure 2. In *feed* there is energy apparent just below the 3kHz upper frequency limit of the display in the F2/F3 region. There is energy in the low frequency F1 region for all three vowels. For the vowel in *part* energy peaks corresponding to F1, F2 and F3 are visible.

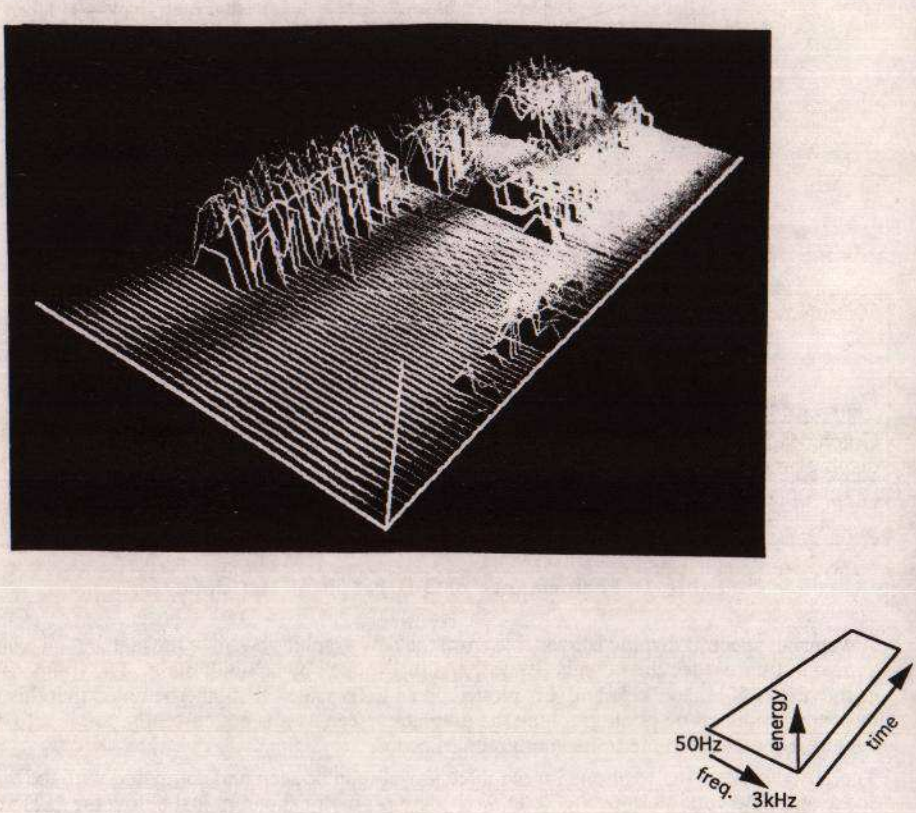

Figure 4: Output waterfall display for the vowels in *feed*, *part* and *food* spoken by an adult male. (NB Frequency plotted on an ERB scale.)

6. CONCLUSIONS

It has been shown that modern parallel computational techniques readily lend themselves to the development of a real-time implementation of the human peripheral hearing system, since the key to the contemporary psychoacoustic models is their parallel nature. The implementation of such models will enable a full exploration of: (a) the functionality of the hearing system itself, (b) the

nature of the acoustic patterns which are important for the perception of speech, and (c) the application of visual displays for the development of speech and singing skills.

The system produces a spectrographic display which is rather crude and we have produced a transputer graphics board which enables a 3-D waterfall spectral display to be produced in real-time. The next stage is to implement a grey scale spectrogram display and to investigate the minimum sweep time which can be implemented.

Such a system could be employed help enable the investigation of acoustic cues in speech based on current knowledge of the operation of the human peripheral hearing system, and the implementation of a new generation of real-time visual displays for use by speech therapists and others involved with the speech and hearing impaired and those involved in professional voice training.

7. ACKNOWLEDGEMENTS

The authors would like to thank Rob Sloan for his photographs of the transputer graphics screen.

8. REFERENCES

Baken, R.J., (1987). 'Clinical measurement of speech and voice', Boston: College-Hill Press.

Moore, B.C.J., (1982). 'An introduction to the psychology of hearing', London: Academic press.

Patterson, R.D., (1976). 'Auditory filter shapes derived with noise stimuli', Journal of the Acoustical Society of America, **59**, 640-654.

Swan, C., Howard, D.M., and Tyrrell, A.M. (1994). 'Real-time simulation of the human peripheral hearing system', Microprocessors and Microsystems, **18**, 215-221.

Tyrrell, A.M., Howard, D.M. and Beasley, N.A., (1992). 'Transputer Model of the Human Peripheral Hearing System', Microprocessing and Microprogramming, **35**, 619-624.