ANNEX A OF THE SVOS FINAL REPORT
(Part A: The Auditory Filterbank)

A PRELIMINARY STUDY OF THE FEASIBILITY OF A HARDWARE VERSION OF THE AUDITORY FILTER BANK

Roy D Patterson

MRC Applied Psychology Unit
15 Chaucer Road
Cambridge CB2 2EF

Peter Rice

Cambridge Electronic Design
Science Park
Milton Road
Cambridge

January 1987

A discussion paper written as an interim report for the "Hardware Filterbank" project and in preparation for the 'standard filterbank' meeting held at the Institute of Hearing Research (IHR), Nottingham, 19 February, 1987.

This paper outlines the research done in the first stage of a project to determine the feasibility of a hardware auditory filterbank that would run in real-time on a set of conventional DSP chips. The first section of the paper outlines the procedure whereby we established the initial filterbank design; that is, the appropriate auditory filter shape, the number and spacing of the filters, and the appropriate bandwidth function. It also considers the available filterbank architectures and filtering algorithms, and concludes that it should now be possible to produce a real-time filterbank for predicting masking in helicopters with the accuracy available in the best auditory filter model. The filterbank would contain 32 bandpass filters covering the frequency range 100 to 5,000 Hz and it would require approximately 5 DSP chips.

The second section of the paper reviews the spectral analysis systems currently used in hearing and speech research to ascertain (a) the extent to which the initial filterbank might satisfy these needs, and (b) how the initial filterbank design might have to be extended to serve the larger community. Nine categories of spectral analysis are considered and it is concluded that the filterbank would have to be extended to cover the frequency range 100 to 10,000 Hz and to include up to 128 filter channels. It would not, however, require any new technology.

INTRODUCTION

Over the past five years the Applied Psychology Unit (APU) in Cambridge, the Institute of Sound and Vibration Research (ISVR) in Southampton, and Cambridge Electronic Design Limited (CED) have provided research support for the Royal Aircraft Establishment (RAE) at Farnborough. The main aim of the collaboration was to devise a practical model of auditory masking that would enable us to predict what is, and what is not, audible in military helicopters, and to devise an optimised set of auditory warnings for use in the helicopter fleet.

Given the time constraints of the project we chose a model of auditory masking that was limited to stationary sounds, and so it was limited to analyses based on considerations of the longterm power spectra associated with the helicopter noise and the auditory warning sounds (Patterson, 1976; Patterson, 1982a). At the heart of this model is a spectral representation of the auditory filter bank which is used (a) to convert the power spectrum of a given background noise into a 'threshold curve' that is, a function showing masked threshold as a function of frequency, and (b) to generate an 'excitation pattern' for each auditory warning to be presented in the noise. The threshold curve can then be compared with the excitation pattern of each signal to determine its appropriate level. With respect to these limited aims, the model proved surprisingly successful (Patterson, Nimmo-Smith, Weber, & Milroy, 1982; Lower & Wheeler, 1985; Rood, 1984; Lower et al., 1986). Note that 'Threshold curves' and 'excitation patterns' are essentially the same function viewed from two different perspectives. Both are the 'convolution' of a longterm power spectrum and the auditory filter; in one case the input is thought of as a masker and in the other it is thought of as a signal. So, a 'threshold curve' is an 'excitation pattern' for a masker and an 'excitation pattern' is a 'threshold curve' for a signal.

As with most simple models, however, it focussed attention on its own inadequacies. In this case it was the inability of the model to follow the temporal fluctuations present in both the helicopter noise and the auditory warnings. As a result, RAE asked the research group whether it might now be time to consider producing a hardware version of the auditory filterbank that would operate in the time domain and so be able to follow the temporal fluctuations in the signals and noise. A brief study of the costs by APU and CED (Cambridge Electronic Design) indicated that a competent 32-channel device with reasonable safety margins on the important parameters could be constructed using digital signal processing chips (DSP) like the TMS320 for about £50,000 per copy. However, since we needed at least four filterbanks for the project, it was decided that we should begin with a feasibility study to see if we could reduce the safety margin on two or more of the variables to bring the cost per copy down by a

factor of four, or more, without compromising the utility of the filterbank.

The feasibility study began in May of 1986. We set out the relevant auditory variables and attempted to define the appropriate range of the variables. At the same time, we investigated the digital filtering algorithms and the filterbank architectures available to us. In addition to this internal research we began a little external research in the form of discussing our plans with other interested parties in order to broaden our perspective and avoid pitfalls that they might have encountered. The purpose of this paper, is to summarise our internal and external research and to provide a discussion document for a 'standard filterbank' meeting to be held at the Institute of Hearing Research (IHR) in Nottingham on February 19, 1987.

I. FILTERBANK CONCEPTS AND THE INITIAL PARAMETER VALUES.

Since the purpose of the project was to produce a time-domain version of an existing spectral-domain filterbank, we began with the existing power spectrum model and considered the issues involved in converting it to a time-domain model.

A. Auditory Filtering

1. Filter Shape

In the power spectrum model of masking, the amplitude characteristic of the auditory filter is approximated by a rounded exponential, or roex, function with from one to four parameters depending on the accuracy of the prediction required. In its simplest form the roex filter shape is a pair of back-to-back exponentials (a La Place distribution) with the top rounded off to make the filter flat at its centre frequency. The roex filter shape is a particularly simple approximation to the empirical filter shapes measured in human observers over the past ten years by a series of investigators beginning with Patterson (1976) and Houtgast (1977). The justification for the roex shape is presented in Patterson et al (1982) where the filter shape is measured as a function of age.

Briefly, a wide range of studies agree that the auditory filter is a bandpass filter with a broad flat top, sharp skirts that delimit the passband, and shallow sloping tails outside the passband. The equivalent rectangular bandwidth (ERB) decreases from about twenty to ten percent of the centre frequency of the filter, as the centre frequency increases from 100 to 5000 Hz. The skirts of the filter that define the passband fall at a rate of about 100 dB per octave at low centre frequencies and 150 dB per octave at high centre frequencies. The dynamic range of the filter increases with signal level indicating that it is affected by the approach to absolute threshold. At moderate levels the dynamic range of the passband is about 35 dB and at high levels it can be as much as 55 dB.

Outside the passband the tails of the filter fall at a much slower rate -- between 20 and 50 dB per octave. The tail on the low-frequency side of the filter is shallower than that on the high-frequency side. In the psychophysical data, asymmetry in the auditory filter is primarily limited to asymmetry in the tails of the filter at high intensities; at moderate levels the passband is approximately symmetric on a linear frequency scale, and even at relatively high levels there is only a small asymmetry in the passband. A review of auditory filter research is presented in Patterson and Moore (1986).

2. The Number and Spacing of the Filters

In the power spectrum model of masking, we use either 512 or 800 frequency bands in the region below 5 kHz (Patterson 1982b; Lower et al., 1986). At any given frequency the threshold curve is the integral of the product of the noise spectrum and the auditory filter. Currently, the excitation pattern is evaluated at each of the points in the spectrum and so it can be viewed as a software filterbank with 512 or 800 filters. The resulting threshold functions are smooth; however it is clear that this analysis employs more filters than is necessary or desirable in a hardware filterbank. There are some psychoacousticians who believe that we might need up to 300 filter channels to accommodate the perceptual discriminations that humans can make. In the current situation however, the question is not what humans can discriminate, but what we need in the first instance for the prediction of masking, and subsequently for automatic speech recognition. At the low end of the range, it appears possible to predict masking with reasonable accuracy using a filterbank with about 30 filters. The justification for this lower limit is that it is the number of filters which when placed "shoulder-to-shoulder" covers the frequency range 50 to 5000 Hz. By "shoulder-to-shoulder" we mean that the attenuation characteristic on the high side of one filter falls below the 3-dB point just as the lower skirt of the next filter rises above the 3-dB point. There is a theorem that indicates that this is the minimum number of filters for an information preserving filter bank, that is, one for which the original signal can be reconstructed by combining the outputs of the filters. This theorem is sometimes used to justify the 30-filter minimum. What seems more likely, however, is that this is the minimum filter density which ensures that all spectral components will fall comfortably within one filter or another. Since it was clear that the number of filters is a prime determinant of the cost, and that one would probably have to double the number of filters to produce a useful improvement in the prediction of masking, we set the initial number of filters at 32.

3. Filter Bandwidth

Recently, Moore and Glasberg (1983a) reviewed all of the papers which reported attempts to measure the auditory filter shape in humans, and they produced a revised critical-band function referred to as the ERB function. For centre frequencies greater than 1.0 kHz the slope of the new function is very similar to that of the traditional critical-band function; however, the actual values are about 20% lower than the traditional values. In the region below 1.0 kHz, the revised function continues to fall, albeit at an ever decreasing rate, whereas the critical-band function levels out. Although the differences between the two bandwidth functions are not often crucial to

masking prediction, the data underlying the ERB function seem preferable, as they are, limited to filter shape measurements. Accordingly, the ERB function is recommended for calculating the bandwidths of the filters and, in turn, the spacing of the filter centre frequencies.

The ERB function shows that the quality factor, Q, improves as centre frequency increases; that is, the bandwidth is not a fixed proportion of the centre frequency. Any attempt to approximate the filter bank with a strictly proportional system leads to filters that are rather too narrow at the lowest and highest frequencies and rather too wide at frequencies in the mid-range. With regard to masking, the errors in prediction associated with a strictly proportional system would probably be tolerable. There did not, however, seem to be any particular advantage in using a strictly proportional system.

4. The Phase Characteristic

The amplitude characteristic of the auditory filter was derived using a power spectrum model of masking which, by definition, is insensitive to phase. There were no data to indicate that the phase characteristic of the auditory filter plays a significant role in auditory masking when the masker is a noise. Consequently, in the first instance, we assumed that any smooth phase characteristic would provide a satisfactory approximation to that of the auditory filter.

B. Filter Design and Filterbank Architecture

A brief study by ISVR using a simple filter-design program showed that a reasonable approximation to the roex filter shape could be provided by an FIR filter with somewhere between 40 and 80 coefficients. In the final version of the filter bank, we want to use our best estimate of the shape of the auditory filter, rather than the roex function which is a compromise between accuracy and mathematical simplicity. The auditory filter is actually a little flatter and broader than the roex filter (Patterson et al., 1982). We also know that the auditory filter shape is a little less flat than the Gaussian filter function (Patterson, 1976). However, it was sufficient at this point to know that one common digital filter design could provide a satisfactory approximation to the class of filters that includes the auditory filter shape, and that it could do so with a tractable number of coefficients.

The brief study also revealed that it was unlikely that a 16-bit integer implementation of the FIR filter could, at one and the same time, provide a good approximation to the rather broad high frequency filters and the rather narrow low-frequency filters, using one fixed sampling rate. The problem is that the FIR technique

distributes its sample points linearly along the frequency axis and so there are far fewer samples in the passband of the low-frequency filters. This is not an insurmountable problem; one can down sample the signal for the low-frequency filters to raise the density of frequency samples. In this case, one would design the filters with the highest centre frequencies first and check each as it was generated to ensure that the density of frequency samples in the passband of the filter was sufficient to provide a good approximation. When the density fell below a critical level one would lowpass filter the input and use a lower sampling rate ('down sampling'). An initial guesstimate indicated that we would not need to down sample more than four or five times which meant that this technique was feasible. The lowpass filtering would introduce phase lags between the different subsets of filters but these between-channel phase shifts only produce a perceptual change when very large, and so it seemed unlikely that this would be an important problem.

With regard to costing, we estimated that 32 filters would require five TMS320 chips and that the cost of the attendant PCB's and filtering would lead to a filterbank cost on the order of £10,000. Accordingly, it seemed reasonable to proceed.

C. The Initial Filterbank Design

In summary, our initial internal research led us to believe that a time-domain filterbank for predicting masking in helicopters was feasible and that its characteristics would be roughly as follows:

(a) The basic shape of the filter would be that specified by psychophysical masking studies on humans and summarised in the roex approximation to the auditory filter shape (Patterson et al., 1982).

(b) The filterbank would contain about 32 bandpass filters set out shoulder-to-shoulder across the frequency region 100 to 5000 Hz. The filter bandwidths and filter spacing would be as suggested by Moore and Glasberg (1983a).

(c) In the first instance, we would use integer FIR filters and down sample to maintain frequency resolution in the passband as the centre frequency of the filter decreased.

(d) The individual filters would have a linear phase characteristic and we would compensate for the between-channel phase shifts introduced by down sampling.

A summary of the characteristics and parameters of this initial filterbank configuration is presented in Table 1.

II. SPECTRAL ANALYSIS SYSTEMS IN HEARING AND SPEECH RESEARCH.

The auditory filterbank sketched in Section I.A. is based on a power spectrum model of auditory masking and is intended to predict auditory masking in noisy environments. As such it represents the interests of psychologists and engineers concerned with basic and applied aspects of signal detection. Before proceeding to build a filterbank, it seemed wise to consider the interests and assumptions of at least two other groups, namely, the auditory physiologists and the speech scientists. Although all three groups are united in the belief that the auditory system begins the processing of sound with a spectral analysis, their descriptions of that analysis and their concern with different assumptions vary widely. Some of the forms of spectral analysis currently in use are set out in Table 3 in accordance with their concern for a) approximating the bandwidth and spacing of auditory filters, and b) the requirement for temporal resolution at the output of the filter bank. In the first row are models that require the minimum temporal resolution and which typically characterise the stimuli by their longterm power spectra (i.e. line spectra). Their attraction is simplicity and they are quite useful for discussions of quasi-stationary sounds like musical notes and aircraft noise. In the second row are models where a moderate amount of temporal resolution is required to divide the acoustic stream into segments as for speech analysis. Finally, in the third row of the table are models which require extreme temporal resolution, either in the form of basilar membrane motion or very shortterm power spectra. As one proceeds down the rows and across the columns of the table the models become more complex and require increasing computational power. The hardware filterbank will be based on one of the models in the righthand cell of the table, that is, a high-resolution, auditory, spectral-analysis system.

A. Auditory Spectral Representations based on Longterm Power Spectra

1. Filterbanks with fixed bandwidth and filter spacing

For most everyday sounds it is quite difficult to know what the sound will be like when presented with the waveform on its own. Much more informative is the longterm power spectrum of the sound represented by the first cell of the table. One can think of the power spectrum as the integrated output of a filterbank wherein the filters are evenly spaced and all have the same bandwidth. It is still quite common to see line spectra used to explain vowel discrimination and the timbre of instruments. The longterm power spectrum of a reasonably stationary sound is also quite a good predictor of auditory masking in the sense that masking is largely a matter of energy concentrations. The power spectrum reveals whether the energy is predominantly in the low, medium, or high frequencies and if the energy of the masker is in the same

region as the energy of the signal, it is more likely to disrupt perception of the signal.

The longterm power spectrum was first suggested as a model of auditory processing by Ohm (1843) and later it formed the basis of the Place theory proposed by Helmholtz (1875, 1912). As scientific theories go, it has had an amazingly long and successful run. In the end, however, it is incorrect and, in some sense, the analyses presented in the remaining eight cells of the table are simply attempts to refine our estimates of the shape, bandwidth, and integration time of the filters. In the auditory system, the filters with low centre frequencies have a narrow bandwidth and a relatively long integration time; whereas, filters with high centre frequencies have relatively wide bandwidths and relatively short integration times.

2. Filterbanks with proportional bandwidth and filter spacing.

The sound level meter with a set of octave band filters or 1/3-octave-band filters, is perhaps the best example of a filter bank in which the bandwidths of the filters and the filter spacing are proportional to the centre frequency of the filter. They were developed to enable one to perform a spectral analysis in the field at the sound source. Strictly speaking, they measure the shortterm power spectrum of the sound, but in practice the data are plotted as graphs which are best described as longterm power spectra. They are used to assess the annoyance of a sound or its potential to produce noise-induced hearing loss, and as such are outside the main scope of the current project. It is worth noting, however, that the sound level meter with a 1/3-octave-band filter set makes a fairly good model of auditory masking for stationary sounds provided the spectrum does not contain abrupt changes in level. The success of the device derives from the fact that the bandwidth is proportional to the centre frequency. Near a sharp spectral edge the device will overestimate threshold because the filters are two to three times wider than those of the auditory system, but in most practical situations this is not an important error since the device fails in the safe direction. The sound level meter is a single channel device whose output is time averaged to provide a measure of the sound intensity and as such is very different from the hardware filterbank that is the object of this project. However, a twenty-channel sound-level meter in which the output of each channel was the instantaneous waveform would be a good first approximation to the hardware filterbank under discussion.

3. Filterbanks with auditory filter shapes, bandwidths, and filter spacing

The last form of spectral analysis based on longterm power spectra is represented by models in which there is a deliberate attempt to produce an auditory spectral analysis. Accordingly, the filters have the shape and bandwidth of human

auditory filters as we know them. The models exists as computer programs for generating threshold curves and excitation patterns. There have been many versions of these auditory power-spectrum models designed to do everything from predicting the pitch shift of the residue (Patterson & Wightman, 1976) and the masking produced by vowels (Moore & Glasberg, 1983b), to predicting masking in aircraft (Patterson, 1982a; Lower et al., 1986)). A series of studies employing listeners in a helicopter simulator indicated that these models can predict auditory masking in this kind of noise background to within the accuracy of the noise measurements, even with the simplest form of auditory filter approximation (Lower et al., 1986). So long as the filter shape and bandwidth are preserved in the hardware filterbank, and provided the number of filters is sufficiently large and they are distributed according to filter bandwidth, the ability of the hardware filterbank to predict auditory masking in stationary backgrounds is assured. The model is inappropriate insofar as it uses the longterm power spectrum of the stimulus rather than the waveform as input.

The psychophysical studies on auditory filter shape have not yet revealed a method for measuring the phase characteristic of the filter. We do know, however, that the auditory system is either insensitive to reasonably slow rates of change of phase, or it chooses to ignore this information. It is for this reason, then, that we choose to extend the power spectrum model to a time-domain model using filters with a linear phase characteristic. The auditory system does not have a linear phase characteristic, but it appears that the auditory phase characteristic is in the same class as the linear phase characteristic, and so it will provide a reasonable first approximation.

As with all spectral analyses, it is important to note the purpose for which these auditory, stationary filterbanks were designed. As a model of human masking, they include the assumption that all of the frequency selective elements in the auditory system, from the basilar membrane through to the ultimate decision making process, can be represented by the auditory filter shape. In contrast many people would expect the hardware filterbank to represent only those frequency selective elements found in the cochlea. The distinction is probably not important for the prediction of masking, either for stationary sounds or time varying sounds, provided the temporal variation is relatively slow. It becomes important if one assumes that there is neural interaction between channels in the form of lateral inhibition to sharpen the frequency selectivity of the system beyond that provided by the cochlea (as in the model of Lyon, 1982, or Gardner & Uppal, 1986). The time-domain extension of the power spectrum model makes the assumption that the effect of lateral inhibition can be handled by using a sharper filter shape and ignoring the lateral inhibition stage. It is possible that a model of this form will ultimately fail when used to predict auditory perception rather than just

masking. It is also possible that it is more efficient to use a smaller number of wider filters and then enhance the contrast of the resulting pattern using lateral inhibition. These, however, are issues that the project may help decide, rather than those which can be currently used in the design. In the hardware filterbank the filter shapes are stored as coefficients in a reloadable memory. Since the power spectrum version requires more, sharper filters, it is likely that the filterbank will be able to accommodate both types of model.

B. Auditory Spectral Representations based on Shortterm Power Spectra

Although few sounds are stationary in the longer term, many important sounds like the notes of music and some speech sounds are reasonably stationary in the short term. Similarly, many noise backgrounds that are not stationary in the longer term can be thought of as stationary in the short term for practical purposes. Auditory models based on the shortterm power spectra of the stimuli are more complicated than those based on the longterm spectra, but they are decidedly simpler than those based on basilar membrane motion or some other multi-channel representation where the waveform in each channel is preserved. Accordingly, there are numerous spectral analysis systems based on the shortterm power spectra of the stimuli.

1. Analysis filters with fixed bandwidths and constant frequency spacing.

Many of the best known spectral analyses are in this category including the speech spectrograph, the sonogram, the original channel vocoder and multi-channel narrow-band spectrum analysers. They provide a moment to moment measure of the sound power in a range of frequency channels and they could be used quite successfully as models of auditory masking. The duration of the moment over which the power is measured can be varied within limits. Unfortunately, like the Fast Fourier transform, on which many of them are built, the filter bandwidth is fixed. As a result, if the correct bandwidth is chosen to match a mid-frequency auditory filter, the device will overestimate threshold at lower centre frequencies and underestimate threshold at higher centre frequencies. The threshold curve produced by these devices in response to white noise is flat, whereas tone threshold for humans rises about 10 dB over the range 0.5 to 4.0 kHz. Furthermore, the integration time is the same for all channels in these systems, whereas that of the auditory system decreases as the centre frequency increases.

2. Analysing filters with proportional bandwidth and spacing

The disadvantages of the standard spectrograph were recognised by the

speech community and as a result we find that later vocoders have filters whose bandwidth and spacing is proportional to the centre frequency. These systems could be used to predict masking for noise environments with time-varying noise backgrounds quite successfully. Furthermore, if the waveforms generated by the individual filters of the proportional vocoder could be digitised and interfaced to a computer, the resulting device would meet many of the requirements of those in the speech community currently investigating hardware filterbanks.

3. Filterbanks with auditory filter shapes, bandwidths and frequency spacing

There are not, currently, any multi-channel hardware filterbanks in which the analysing filters have the shape, bandwidth and frequency spacing of those in the auditory system. There is, however, one device that comes quite close and that is the spectral analysis performed by the analysis filters of the JSRU channel vocoder (Holmes, 1980). The filterbank has 19 channels with centre frequencies ranging from 240 to 3,750 Hz. The bandwidths of the filters rise from 120 Hz to 300 Hz and the distribution is intended to mimic the older critical-band function of Zwicker (1961). In speech signals, the level variation from channel to channel is not normally greater than 15 dB and so the dynamic range of the passband of the filter is not critical. In the JSRU vocoder, second-order, Butterworth bandpass filters are used, and they were found to be quite satisfactory for many speech applications.

If the outputs of the filters in the vocoder were presented as a 19 channel bar graph, the device would prove quite satisfactory for predicting masking in stationary and quasi-stationary noise backgrounds like those found, for example, on the flightdecks of civil aircraft. In the case of helicopter noise, however, where there are strong line components contributed by the gears and rotors, the second-order Butterworth filters would be too wide for the device to serve as a model of masking. Furthermore, the bandwidths of the filters are about 20% greater than those of the corresponding auditory filters.

C. Auditory Spectral Representations with Auditory Temporal Resolution

1. Filterbanks with fixed bandwidth and filter spacing

There are now several forms of very high resolution spectrogram available; that is, devices that plot or display a sequence of log power spectra where the temporal window is so short as to allow resolution of temporal events within individual cycles of the pitch of speech. The pitch of the female voice and the pitch of childrens voices can be as high as 400 Hz, thus necessitating temporal resolution of events on the order of a millisecond or less. The resulting spectrograms are useful in as much as they reveal,

not only the tracks of the individual formants, but also the shape of the formant within the pitch period. Despite the improved temporal resolution, however, they have the same problem as their predecessors when it comes to the prediction of masking; the fixed bandwidth and filter spacing associated with this spectral analysis means that it will overestimate threshold at low frequencies and underestimate threshold at high frequencies. Accordingly, the high-resolution spectrogram is unlikely to provide a suitable basis for a hardware filterbank.

2. Filterbanks with proportional bandwidth and filter spacing

There are almost no examples of this category of spectral analysis. The one exception would appear to be the filterbank proposed by Lyon (1982). The purpose of this, and subsequent versions of the model, is to provide a frontend spectral processor for automatic speech recognition based on passive basilar membrane motion as characterised in physiological models prior to about 1980. In these models the lowpass part of the auditory filter characteristic is sharp (over 100 dB per octave), but the highpass part of the filter characteristic is very shallow (on the order of 6 dB per octave). Such a system is very unlikely to provide the basis for the hardware filterbank because it greatly overestimates the upward spread of masking observed in psychophysical experiments at moderate to loud intensities. Indeed, it even overestimates the upward spread of masking observed in helicopters (Lower et al., 1986) where the intensity of the rotor component can exceed 120 dB. Furthermore, we now know that the lowpass part of the filter characteristic associated with basilar membrane motion is much sharper than that provided by the passive membrane model underlying Lyon's (1982) simulation.

3. Filterbanks with Auditory Filter Shapes, Bandwidth, and Filter Spacing

There are a very large number of examples in this final category where the aim is to provide a fairly complete simulation of the frequency resolution observed in humans. Unfortunately, the vast majority are physiological models intended to characterise the frequency resolution observed in the cochlea either on, or just after, the basilar membrane, and these models are constructed without regard to complexity or computational load. Such models tend to concentrate on the selectivity of individual channels, and do not consider the more practical questions involved in constructing a filterbank -- in particular, the tradeoff between the accuracy of the filter representation and its computational load, the density of filters required for an adequate simulation of speech processing, and the efficiency of the digital filters used to approximate the observed selectivity.

The problem of specifying a filterbank that is both reasonably accurate and reasonably efficient has been addressed, but more often by speech scientists than either physiologists or psychophysicists. Examples include the filterbanks suggested by Seneff (1984), Lyon & Dyer (1986), Beet, Moore & Tomlinson (1986), Cooke (1986) and Gardner and Uppal (1986). Unfortunately, the majority of these filterbanks are based on passive models of basilar membrane motion, and as such they predict far too much upward spread of masking.

D. Summary of Existing Filterbank Designs

It is not possible, at this point, to determine whether any one of the existing spectral analysis systems in the high-resolution, auditory category would provide a significantly better basis for a hardware filterbank than the others. Indeed, the purpose of this research programme is to make that decision. Nevertheless, the review at this point including discussions with physiologists and speech scientists suggests that the best starting point would be a modified version of the initial configuration presented at the end of the first section of this report and summarised in Table 1. The primary modifications are to the filter density and the maximum sampling rate, with the speech community requiring as many as 128 filters covering the range 50 to 10,000 Hz, and a maximum sampling rate of 25 kHz necessitated by the desire to have filters with centre frequencies as high as 10 kHz. The characteristics of this generalised filterbank are presented in Table 2. It should be noted that the configurations presented in the initial and generalised configurations are quite similar (Tables 1 and 2). They both employ a roex filter shape, with ERB-function bandwidths and filter spacing. Furthermore the architecture of the hardware and software proposed for the filterbank is essentially the same; it is simply that the generalised configuration would be two to four times larger than the initial configuration. This suggests that we might proceed with the aim of producing a hardware filterbank that meets the initial configuration and which will probably suffice for predicting masking in helicopters. Then, an extended version of the filterbank could be instructed should the speech community require it.

ACKNOWLEDGEMENTS

This work has been carried out with the support of Procurement Executive, Ministry of Defence.

REFERENCES

Beet, S.W., Moore, R.K., & Tomlinson, M.J. (1986). Auditory modelling for automatic speech recognition. Proceedings of the Institute of Acoustics: Speech and Hearing, Vol. 8, Part 7, 571-580.

Cooke, M.P. (1986). Towards an early symbolic representation of speech based on auditory modelling. Proceedings of the Institute of Acoustics: Speech and Hearing, Vol. 8, Part 7, 563-570.

Gardner, R.B., & Uppal, M.K. (1986). A peripheral auditory model for speech processing. Proceedings of the Institute of Acoustics: Speech and Hearing, Vol. 8, Part 7, 555-562.

Helmholtz, H.L.F., von (1875, 1912). On the Sensations of Tone. English translation of 4th edition by A.J. Ellis (Longmans, Green and Co., London, 1912).

Holmes, J.N. (1980) The JSRU channel vocoder. IEEE Proceedings, Vol. 27, Pt. F., No. 1, 53-60.

Houtgast, T. (1977). Auditory-filter characteristics derived from direct-masking data and pulsation-threshold data with a rippled noise-masker. Journal of the Acoustical Society of America, 62, 409-415.

Lower, M.C., & Wheeler, P.D. (1985). Specifying the sound levels for auditory warnigs in noisy environments. In I.D. Brown, et al., (Eds.) Ergonomics International 85, Taylor & Francis, 226-228.

Lower, M.C., Patterson, R.D., Rood, G., Edworthy, J., Shailer, M.J., Milroy, R., Chillery, J., & Wheeler, P.D. (1986). The design and production of auditory warnings for helicopters 1: the Sea King. Institute of Sound and Vibration Research Report AC527A.

Lyon, R.F. (1982). A computational model of filtering, detection, and compression in the cochlea. Proceedings, IEEE ICASSP, Paris, 1282-1285.

Lyon, R.F., & Dyer, L. (1986). Experiments with a computational model of the cochlea. Proceedings, IEEE ICASSP, Tokyo, 1975-1978.

Moore, B.C.J., & Glasberg, B.R. (1983a). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns", Journal of the Acoustical Society of America, 74, 750-753.

Moore, B.C.J., & Glasberg, B.R. (1983b). Masking patterns for synthetic vowels in simultaneous and forward masking. Journal of the Acoustical Society of America, 73, 906-917.

Ohm, G.S. (1843). Uber die Definition des Tones, nebst daran geknüpfter Theorie der Sirene und ähnlicher tonbildender Vorrichtungen. Annals Physical Chemistry, 59, 513-565.

Patterson, R.D. (1976). Auditory filter shapes derived with noise stimuli. Journal of the Acoustical Society of America, 67, 229-245.

Patterson, R.D. (1982a). "Guidelines for auditory warning systems on Civil Aircraft", Civil Aviation Authority, Paper 82017, London.

Patterson, R.D. (1982b) "Review of the auditory warning system proposed in ARINC 726." Civil Aviation Authority Contract Report 7D/S/0260/3.

Patterson, R.D., & Moore, B.C.J. (1986). Auditory filters and excitation patterns as representations of frequency resolution. In B.C.J. Moore (Ed.) Frequency Selectivity in Hearing. Academic: London, 123-177.

Patterson, R.D., & Wightman, F.L. (1976). Residue pitch as a function of component spacing. Journal of the Acoustical Society of America, 59, 1450-1459.

Patterson, R.D. Nimmo-Smith, I. Weber, D.L., & Milroy, R. (1982). The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold. Journal of the Acoustical Society of America, 72, 1788-1803.

Rood, G.M. (1984). Predictions of auditory masking in helicopter noise. Tenth European Rotorcraft Forum, August 38-31, 1984 - The Hague, The Netherlands, Paper Nr. 20.

Seneff, S. (1984). Pitch and spectral estimation of speech based on auditory synchrony model. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, San Diego, March 1984, paper 36.2, vol. 3.

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands. (Frequenzgruppen), Journal of the Acoustical Society of America, 33, 248.

Table 1 Hardware Filterbank Feasibility Study (APU/RAE/ISVR/CED)
Initial Configuration, May 1986

PURPOSE: Simulation of peripheral auditory processing to predict masking in aircraft with non-stationary noise environments

CATEGORY: auditory filterbank, high temporal resolution

FILTER CHARACTERISTICS:

Shape:	auditory, roex(p): $W(g) = (1+pg)\exp(-pg)$, $g = \text{abs}[(fc-f)/fc]$
Bandwidth:	ERB function (Moore & Glasberg, 1983); 130 Hz at 1.0 kHz
Spacing:	auditory, shoulder-to-shoulder, ERB function
Number:	32 filters covering 100 - 5000 Hz

IMPLEMENTATION

Hardware: a set of DSP chips mounted on PCB's in a box

- analogue or digital input; single channel
- digital multi-channel output
- analogue output, one channel at a time
- anti-aliasing filter on analogue input and output

Software:

- FIR filters to simplify design and ensure stability
- integer implementation to achieve the necessary speed
- linear phase characteristic
- 12 kHz sampling rate
- downsampling to reduce the number of coefficients required to implement the narrow low-frequency filters

Table 2 Hardware Filterbank Feasibility Study (APU/RAE/ISVR/CED)
General Configuration: February, 1987

PURPOSE: Simulation of peripheral auditory processing to predict masking in aircraft with non-stationary noise environments.

CATEGORY: auditory filterbank, high temporal resolution

FILTER CHARACTERISTICS:

Shape:	auditory, roex (p): $W(g) = (1+pg)\exp(-pg)$, $g = \text{abs}[(fc-f)/fc]$
Bandwidth:	ERB function (Moore & Glasberg, 1983), 130 Hz at 1.0 kHz
Spacing:	auditory, ERB function, four filters per ERB (i.e. 4 times the shoulder-to-shoulder density)
Number:	128 filters covering 50 - 10,000 Hz

IMPLEMENTATION

Hardware: a set of flating-point DSP chips mounted on PCB's

- analogue or digital input; single channel
- digital multi-channel output
- analogue output, one channel at a time
- anti-aliasing filter on analogue input and output

Software:

- FIR filters to simplify design and ensure stability
- floating point implementation for accuracy and to provide for recursive filter designs later
- linear phase characteristic
- 25 kHz sampling rate
- downsampling to reduce the number of coefficients required to implement the narrow low-frequency filters

TABLE 3 SPECTRAL ANALYSIS SYSTEMS FOR HEARING AND SPEECH RESEARCH

Temporal Resolution (sample dur)	Filter Bandwidth and Filter Spacing		
	Fixed	Proportional	Auditory
Minimal (1000 ms)	Longterm power spectrum (line spectrum)	Sound level meter with 1/3 oct filter set	Masked threshold curve Excitation pattern
Moderate (10 ms)	Spectrograph Sonograph Original channel vocoder Multi-channel narrow-band spectrum analyser	Proportional, channel vocoder	JSRU vocoder Auditory multi-channel spectrum analyser
High (0.1ms)	High-resolution spectrogram	Lyon (1982) filterbank	Auditory hardware filterbank Lyon (1984)/Beet et al (1986) Seneff (1984) Gardner & Uppal (1986) Cooke (1986) Patterson (1987)