Proceedings of The Institute of Acoustics

AUDITORY MODELLING FOR AUTOMATIC SPEECH RECOGNITION

S.W. Beet, R.K. Moore and M.J. Tomlinson

Speech Research Unit, R.S.R.E., Malvern

INTRODUCTION

Previous work at R.S.R.E. has shown that the resolution offered by conventional filter-bank analysis is neither sufficient for accurate temporal modelling of speech signals nor for reliable separation of speech from competing signals and noise. In order to further investigate these problems, a software model of the human peripheral auditory system has been implemented, based on that described by R.F. Lyon [1, 2, 3]. Lyon's model, however, produces data at an extremely high rate and it is consequently unsuitable for immediate use by current recognition algorithms.

This paper describes a *reduced auditory representation* (R.A.R.), which preserves the main characteristics observed in the output of the auditory model, including the ability to resolve fine temporal and spectral detail (without excessive disruption by individual pitch pulses), but which produces data at a more acceptable rate.

THE LYON AUDITORY MODEL

Many structures have been suggested for simulating the observed behaviour of the human auditory system, but one of the most comprehensive versions to retain computational tractability is that developed by Richard Lyon at Schlumberger Palo Alto Research. This model approximates the behaviour of the inner ear, while effects due to the outer and middle ears are largely neglected. The latter are omitted from the model because they merely impose a fixed, albeit position and environment dependent, spectral distortion on any signal and it is known that human perception is relatively insensitive to such distortions. Furthermore, since the input to the auditory model is derived from a single microphone subjected to a complex sound field within a generally unknown environment, it would not be practical to include an accurate model of these parts of the system; the only way of incorporating such effects, short of measuring the whole sound field with an array of microphones, would be to mount a single microphone inside a physical model of an ear.

Lyon's model has been described at various stages of its development and there have often been significant changes between each publication. However, what follows is a description of the main features of the most recent version (unless stated otherwise).

The basilar membrane model

In his earlier papers, Lyon modelled the basilar membrane using a cascade/parallel filter bank, preceded by a differentiator [1]. The cascade of notch filters simulates the propagation of a pressure wave along a simplified (single-dimensional and reflectionless) model of the basilar membrane, while initial differentiation, together with the parallel resonator sections, model the transformation from pressure to basilar membrane displacement. A further anti-resonance is introduced into each output at half the resonator's centre frequency in order to model the measured phase response of real auditory systems more accurately.

Lyon later simplified this part of the model (by modifying the notch filters and omitting the resonators [3]) so as to reduce the computational load, but the R.S.R.E. implementation has retained the original form to ensure a more realistic approximation to the physical system.

Dynamic range compression

Before the basilar membrane displacement is converted to neural firings, several effects observed in the human auditory system are simulated by three layers of dynamic range compression, each with a different rate of adaptation. The effects thus simulated are the stapedial reflex, lateral suppression, firing rate adaptation and the exceptionally wide dynamic range of input signals which can be handled by the system. The algorithm which Lyon originally suggested was found to be potentially unstable, so he later developed a stable, recursive form which adjusts the gain of each channel so as to compress the overall dynamic range while maintaining local contrasts [2, 3]. This has a speed of adaptation determined by a single parameter chosen to match physiological observations. Additional factors are introduced into the algorithm to ensure a smooth gain profile across the membrane, reproducing the phenomenon of lateral suppression. The strength of this suppression is determined by two parameters, one governing the suppression by signals in channels closer to the apical end of the membrane, while the other governs suppression by those closer to the basal end.

The hair cell/neuron model

The ultimate firing of the neurons is attributable to the flow of current through the inner hair cells [4] and it is thought that this current flow is modulated by the deformation of these cells causing their resistivity to change (there is an approximately constant potential difference between the fluid on either side of the basilar membrane, so any change in resistance will produce a modulation of the current flow through that cell). The current flow is integrated (to simulate the accumulation of ions within the cell) until the 'neuron' fires, producing an output event and removing the 'accumulation of ions' by resetting the integrator's stored potential.

This level of the model has a binary output (whereas the information present in the earlier stages is in the form of a continuously varying function of time) so the information is coded in a purely temporal form. This is the first stage in the model which cannot be analytically inverted, but in the human auditory system there are a very large number of afferent neurons, allowing the original signal to be characterised unambiguously. It is impractical to compute the earlier stages of the model at the required number of points along the basilar membrane, but this restriction can be partially overcome by attaching more than one 'neuron' to the output of each 'hair cell'. Each neuron is set to have a slightly different (randomly selected) set of parameters so that the firing events will occur at different times. Thus, although some of the ability to discriminate between closely spaced frequencies may have been lost, the information at the preceding level will have been adequately coded.

THE REDUCED AUDITORY REPRESENTATION

Lyon's model has been used to analyse various signals including a limited amount of natural speech, an example of which is shown in figure 1. This is taken from the utterance "two", the output of each major step in the process being shown separately. The input waveform is plotted across the top of each band and it is clear that the time taken for the model to respond to any onset is very short and that the temporal resolution is thus very fine. Several features of the signal are made explicit by this processing and some aspects are visible in more than one form. For example, the first formant is visible (primarily during the later, voiced, part of the display) both as a disruption in the coninuity of the pattern about a third of the way up each display and as the temporal modulation of neuron firings over a large portion of the model (at those points where no other signal dominates).

Reducing the data rate

It is assumed that the data produced by the auditory model contains the 'relevant' aspects of any speech signal in a more explicit form than conventional analyses and would ultimately lead to improved recognition performance. However, current computational constraints prevent the direct use of the vast amount of data produced by such an analysis (about 30 Mbits per second). Therefore some form of down-sampling is required and this should be chosen so as not to disrupt the structure introduced by the preceding analysis. In order to achieve this *reduced auditory representation*, one has to define exactly what constitutes 'structure' and at this point some gross simplifications are made so as to ensure that the analysis process forms a reasonably cohesive whole and is mathematically concise:

1) Individual neuron firings are not calculated; the most reliable indications of a signal's properties are assumed to be the result of multiple neural events and therefore global statistics of those events can be used to characterise the signal.

2) The observed ability of a neuron to fire in synchrony with a subharmonic of a signal is ignored. This may degrade the analysis' ability to detect a common harmonic structure in different channels, but allows considerable simplification.

3) The main cues to the identity of the various components in a composite waveform are assumed to be the neural firing density at each point along the membrane and any periodicity in those firings.

4) Although the phase relationship between adjacent points on the membrane may be used by the human auditory system, it is not likely to be critical to the analysis of speech since such signals are perceived clearly even when subjected to quite severe phase distortion. Such relationships are, therefore, ignored.

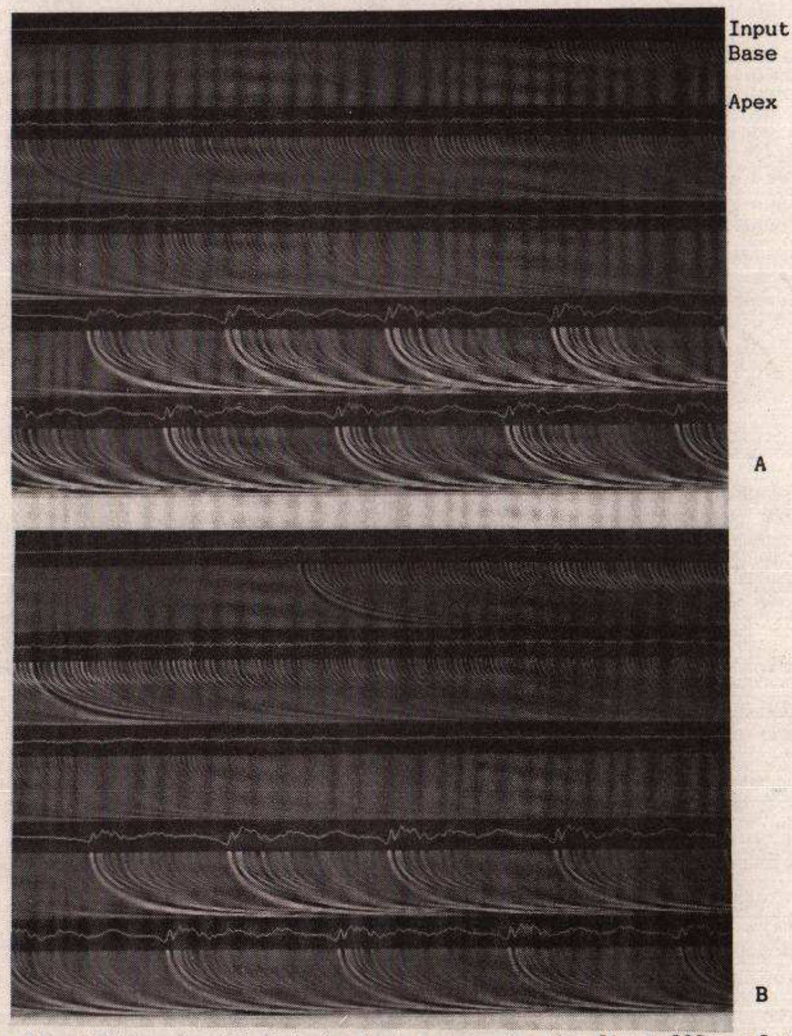

Figure 1: Auditory model outputs in response to the first 200ms of the utterance "two". Positive displacement is denoted by white areas, negative by black.

A: Basilar membrane displacement
B: After dynamic range compression

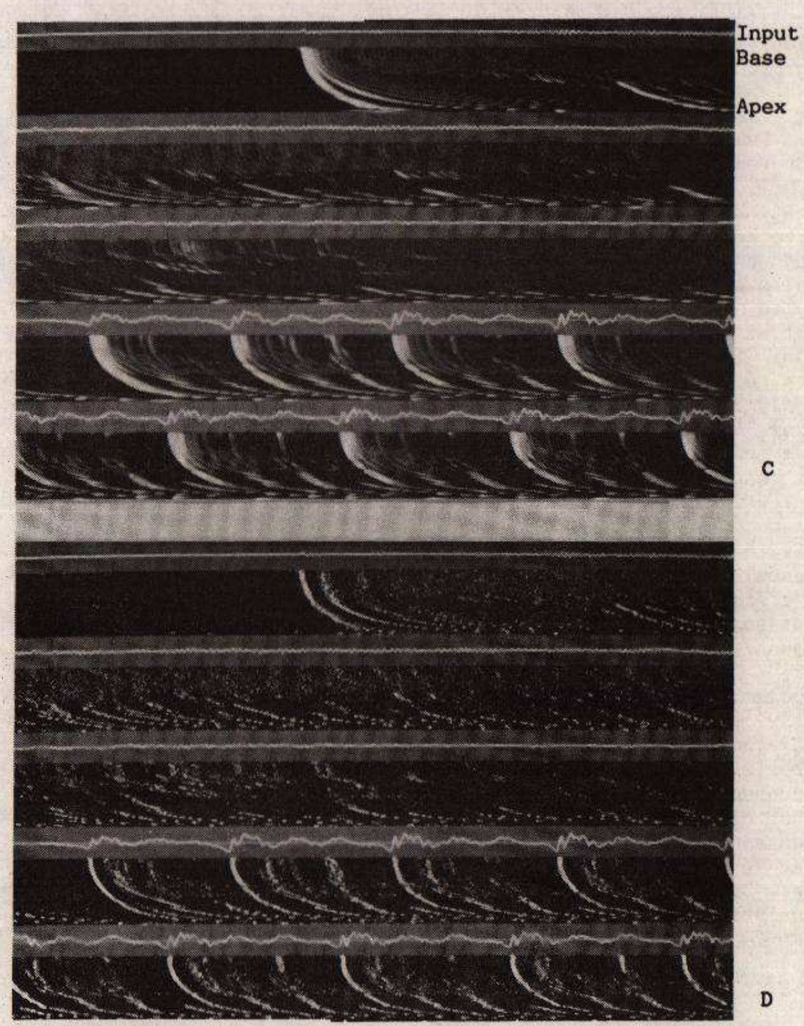

Figure 1 (Continued)

C: Hair cell current flow
D: Neuron firings

These simplifications are all naïve and it is quite possible that significant information will be lost as a result. However, some such assumptions must be made if a practical method is to be developed and these allow an analysis to be performed which is superior to conventional filter-bank analyses while retaining close links with some of the more traditional concepts in speech analysis. Furthermore, it is expected that some of the simplifications will be changed or removed altogether when more experience has been gained.

Simplifications 1 and 2 allow the actual neuron firings to be ignored and the analysis to be based purely on the preceding levels of the model (neural firing rates and any periodicity therein can be estimated without calculating individual events). 3 and 4 suggest that the functions which are to be down-sampled need only be capable of yielding an estimate of the amplitude of the signal at each point on the membrane together with some corresponding measure of frequency. However, the emphasis which the neuron firings would have given to onsets and spectral discontinuities should be retained. This is most readily achieved by producing parameter estimates at each instant and then calculating a weighted mean of the resultant values based on the expected neural response. If this approach is used, continuity from one time frame to the next can be ensured by allowing the frames to overlap and introducing an additional weighting factor, typically consisting of a raised-cosine window.

As regards the detection process, it seems likely that the human auditory system uses the deformation of the hair cell simply because that mechanism is the most efficient physiological solution. This stage of the auditory model does not appear to reveal any significant features in the compressed membrane displacement and so a superior detection process can be used.

At this point, it is further assumed that the output of each filter will essentially consist of a summation of a number of modulated sinusoids (typically those harmonics of any periodic signal which are contained within the filter's pass-band). This is reasonable, even when the model is being fed with broad-band noise, because the basilar membrane will introduce large amounts of correlation into the signal. One of the simplest ways of measuring the instantaneous amplitude of a single modulated sinusoid is to produce the equivalent analytic signal (via the Hilbert transform) and evaluate its magnitude. Similarly, the frequency can be measured by dividing the derivative by the analytic signal itself:

$$a(t) = \left| S_A(t) \right| \quad ; \quad \omega(t) = \mathrm{Im}\left(\frac{1}{S_A(t)} \frac{dS_A(t)}{dt} \right) \tag{1}$$

where $a(t)$ is the amplitude, $\omega(t)$ is the angular frequency and $S_A(t)$ is the analytic signal derived from the output of each filter, all being functions of time, t.

These functions yield meaningful values when the signal under consideration can be considered to be a single modulated sinusoid. However, if it can only be realistically represented as a composite waveform, the values produced can exhibit extreme local variability and will be unsuitable for down-sampling. Therefore it is proposed that the following functions should be evaluated and down-sampled:

$$a'(t) = \left| S_A(t) \right|^2 \quad ; \qquad d'(t) = \left| \frac{dS_A(t)}{dt} \right|^2 \qquad (2)$$

If these expressions are estimated from a large number of consecutive samples, yielding $\hat{a}(t)$ and $\hat{d}(t)$, then $\hat{a}(t)$ will be a weighted mean square of the amplitudes of all the components contained within the analysis frame. Similarly, a weighted mean square of the frequencies of the components can be produced by evaluating

$$\hat{\omega}(t)^2 = \frac{\hat{d}(t)}{\hat{a}(t)} \qquad (3)$$

When calculating $\hat{a}(t)$ and $\hat{d}(t)$, a *neural weighting function* should be used which will make those estimates reflect the emphasis placed on each instant by the number of neurons which would have responded at that time in an 'idealised' auditory system. It is assumed that such a system would contain an infinite number of neurons and that the number firing in response to any onset would be proportional to the amplitude of the signal at that time. It is further assumed that an ideal form of dynamic range compression would cause the number of neurons responding to a steady signal to decay exponentially with time. Now the expressions for $a'(t)$ and $d'(t)$ are inherently weighted in favour of large amplitude signals, so provided that dynamic range compression is performed before the parameters are evaluated, no additional weighting factor is required (except for that due to the sliding window mentioned earlier). However, dynamic range compression can affect the spectral purity of a signal (by temporal modulation of the original function) so it would be better if it was replaced by a further weighting factor which would reduce the effect of any stationary components. Such a factor should counteract that due to the signal's amplitude as a steady state is approached, the obvious choice being the reciprocal of a smoothed estimate of the signal amplitude based on its previous values and those of the more apical sections of the filter (to simulate temporal and upward masking).

Thus all the parts of the original model, except the initial basilar membrane filtering, have been replaced by idealised forms and the resulting parameter estimates may well reflect those which the ear would produce if it were not limited by physiological considerations. The filtering has not been altered because the original model has provided a computationally efficient method for achieving the required result.

PROPERTIES OF THE R.A.R. ANALYSIS

The *reduced auditory representation* of a segment of speech is shown in figure 2. Again, the input waveform is shown across the top of the picture, with the amplitude and frequency parameters displayed in the two bands below. The amplitude display covers a range of about 60dB, while the frequency values are normalised for each channel so that white indicates the presence of a dominant signal near the top edge of each filter's pass band and black indicates one with less than a third of that frequency.

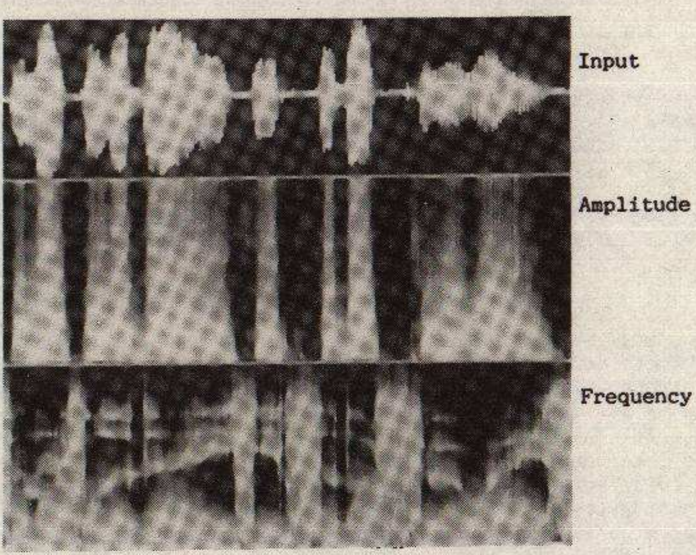

Figure 2: R.A.R. of "An apple a day keeps the doctor away."

It can be seen from this picture that the frequency information provides the clearer indication of the location of spectral peaks, but the amplitude must also be used to differentiate between some classes of signal. In addition, information about the shape of any spectral peaks can be obtained by comparison of the two forms of output. For example a broad spectral peak will produce bright areas on both displays at about the same point, while a pure tone will produce a high frequency estimate nearer to the apex than the corresponding peak in the amplitude estimate. Temporal onsets can also be characterised in a similar way (abrupt onsets show a high frequency estimate before the amplitude has risen significantly).

An additional feature of the analysis which is not so obvious from the picture is due to the neural weighting function used to produce the parameter estimates. This acts as a bias which emphasises those components in each filter's output which are unexpected (in terms of its previous outputs and those of its apical neighbours). This aids in the resolution of proximate but distinct signals and should be beneficial to the discrimination of complex sounds produced by independent sources.

CONCLUSION

The R.A.R. analysis produces clear cues to the location of formants, but it also has several new and interesting properties which should provide improved speech recognition performance. The main features are the absence of any amplitude dependence in the frequency estimates (making the resultant patterns relatively immune to spectral tilt and amplitude variation) and the ability to resolve fine temporal and spectral detail. These features will be especially important for speech recognition if the signal has been corrupted by spectral distortion or the addition of extraneous signals, as is often the case. Furthermore, the physiological base from which the analysis has been derived leads us to expect that the features which it reveals should be more perceptually relevant and therefore better suited to application in a high-performance speech recogniser.

REFERENCES

1 R.F. Lyon, 'A computational model of filtering, detection and compression in the cochlea', IEEE ICASSP-82, 1282-5 (1982)

2 R.F. Lyon and N. Lauritzen, 'Processing speech with the Multi-Serial Signal Processor', IEEE ICASSP-85, 981-5 (1985)

3 R.F. Lyon and L. Dyer, 'Experiments with a computational model of the cochlea', IEEE ICASSP-86, 37.6.1-4 (1986)

4 J.B. Allen, 'A hair cell model of neural response' in 'Mechanics of Hearing', 193-202, Eds. E. de Boer and M.A. Viergever, Martinus Nijhoff (1983)