Proceedings of the Institute of Acoustics

# INTERFACING AN AUDITORY MODEL TO A PARAMETRIC SPEECH RECOGNISER

S. W. Beet and I. R. Gransden

University of Sheffield,
Department of Electronic and Electrical Engineering (Electronic Systems Group),
Mappin Street, Sheffield, S1 3JD.

## 1. INTRODUCTION

This paper describes the results of some simple vowel-discrimination experiments based on isolated frames of data. That data was produced from short segments taken from the TIMIT database, and took the form of the new, linear frequency-scale, version of the Reduced Auditory Representation (RAR) [1]. The recognition was performed using standard linear discriminant techniques, albeit with a rather lower dimensionality than normal because of the reduced number of classes which were to be identified.

## 2. THE 'ORIGINAL' RAR

### 2.1 The Basic Principle

The RAR is a method for analysing acoustic signals (speech in particular), which is based on a functional model of the peripheral auditory system. It includes models of adaptation, masking and loudness compression, and exhibits most of the phenomena observed in physiological experiments to a reasonable degree of accuracy.

Although the RAR has evolved over many years [2,3,4,1], it has consistently been founded on the premise that it is not solely the mean neural firing rate which characterises sounds in the auditory nerve. Other features are just as likely to be important to human perception (and in some cases, more so). The RAR therefore provides four parameters for each point along the cochlear partition and for each point in time: a mean firing rate (related to signal intensity and encoded on a logarithmic scale), an adaptation factor (also logarithmically encoded), a dominant frequency and a phase delay between adjacent channels. The first two of these are determined by the static and dynamic aspects of the signal amplitude, respectively, while the last two are functions of the component frequencies.

### 2.2 The Need for Phase Information

Of particular interest, both for signal characterisation and for source separation, are the synchrony between neurons responding to a common signal component, and that between those responding to different components originating from the same source. These 'synchrony' factors are essentially functions of the phase structure of the basilar membrane displacement, and so are not present in long-term neural firing rate data. They can, however, be characterised by phase derivatives, which are (fortunately) slow to change in most cases. This is the approach taken in the RAR analysis.

By making estimates of the phase derivative with respect to cochlear position, as well as that with respect to time, a more complete description of local synchrony is obtained. Global synchrony is more problematical though, since attempts to identify synchrony between arbitrary combinations of signal components tend to lead to a 'data explosion' with too many possible ways of combining those components. This problem has not yet been addressed in the RAR.

### 2.3 Back to Basics

The other underlying principle behind the RAR analysis is that it is kept mathematically simple (and hence easy to understand, efficient to compute on DSP hardware and with predictable behaviour, even when presented with complicated signals such as speech). In practice, this means that all four parameters are calculated as weighted averages of instantaneous estimates of the respective values. Thus each parameter is the result of integrating two functions over a common window, and then dividing one by the other.

The nature of the weighting function (the denominator in the preceding description) is chosen so as to produce 'correct' results, assuming that the window is large enough, while emphasising the high-energy sections within the signal. In this way, the parameter estimates tend to reflect whichever of the components is 'dominant' at any given point. The RAR is therefore less blurred than a more conventional analysis might be when presented with a composite signal.

## 3. RECOGNISING SPEECH FROM ITS RAR

To avoid the problems of pitch harmonics and/or pitch pulses disrupting the RAR data and causing subsequent misrecognition, a 'position-tolerant distance measure' was introduced in [3]. This, together with the long duration of the RAR's temporal integration window, produced extremely encouraging results. The next step in this line of investigation should have been to extend the 'position-tolerant' concept to include the temporal dimension and avoid the integration process altogether. However the method, as presented in [3] and [4], was based on the idea that the recogniser was performing a pattern-matching task, rather than any form of parametric modelling. This meant that the new distance measure was inapplicable to continuous-distribution hidden Markov models (HMMs). As it turned out, it was also inappropriate for discrete-distribution HMMs, because of the high dimensionality of the RAR data and the consequent difficulty of performing sensible vector quantisation[1].

One possible solution to these problems might be to use the 'position-tolerant distance measure' in a 'fuzzy vector quantiser'. This approach is currently being investigated, although the results described here were obtained using a more obvious and somewhat less interesting technique.

[1] If the codebook were too small, or the centres placed in inappropriate positions, information would be lost and recognition performance degraded. Conversely, if it were large enough to ensure that there were centres near every conceivable region of importance, then the amount of training data needed to build statistically reliable probability estimates, would be vast.

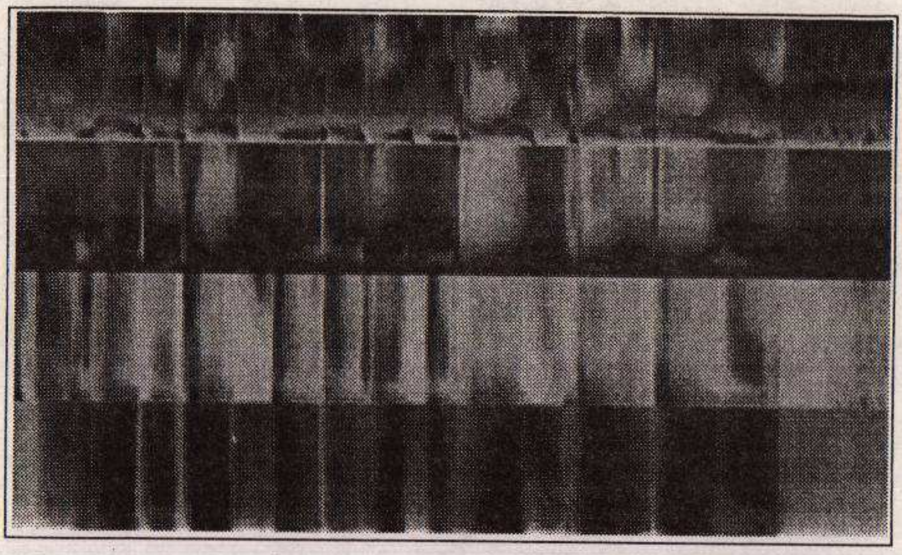

Figure 1: RAR of a typical TIMIT utterance optimised for resolution of events with characteristic duration of 12.5 ms and bandwidth of 400 Hz. Each band represents one RAR parameter. From the top down, these are the delay, frequency, adaptation and intensity, respectively.

## 4. MODIFYING THE RAR

### 4.1 Spatio-Temporal Integration

To overcome the problems associated with changes in pitch, some form of spatial integration was required (in addition to the temporal integration already inherent in the calculation of the RAR parameters). This spatial integration has not been directed by theories of speech perception (although this has been investigated in some detail by others at Sheffield [5]). Instead, a purely pragmatic approach has been taken: it has been assumed that the highest pitch likely to be analysed is 400 Hz, so the outputs of all channels with centre frequencies within about 400 Hz of any chosen reference point are weighted and summed as part of the existing (previously solely temporal) averaging process. Similarly, the maximum expected pitch period of 12.5 ms is used to define the extent of the temporal averaging. An example of an RAR designed to give these respective degrees of temporal and spectral resolution is shown in figure 1.

This description is actually somewhat oversimplified, since the form of the weighting in both spatial and temporal dimensions is chosen quite carefully, so as to make the most of the

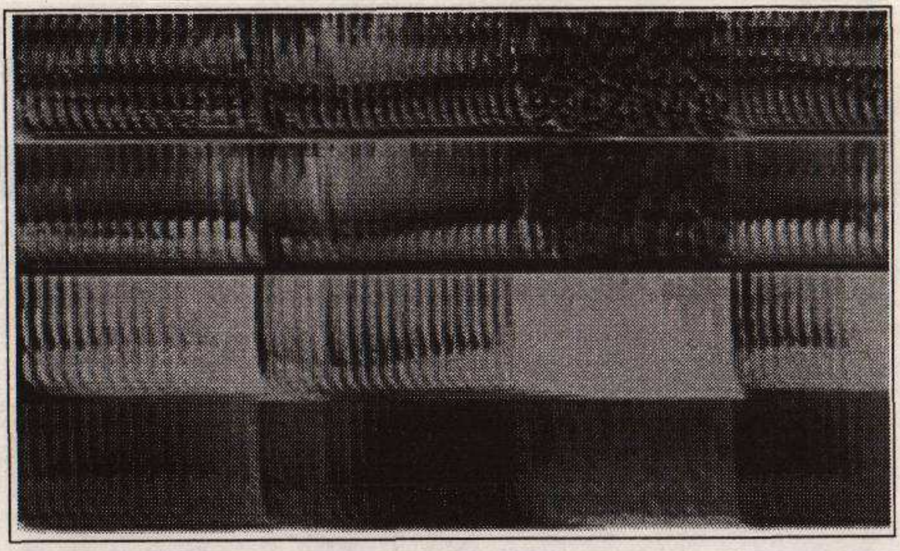

Figure 2: RAR of a syllable extracted from a typical TIMIT utterance optimised for resolution of events with characteristic duration of 1.25 ms and bandwidth of 40 Hz. Each band represents one RAR parameter. From the top down, these are the delay, frequency, adaptation and intensity, respectively.

resolution available at the various channel outputs [1].

### 4.2 Information Loss

Although the aim of this process is to remove any evidence of pitch variation, it should be noted that it can have other advantages if the data is, say, intended for visual representation, rather than automatic speech recognition. In such cases, it need not result in information loss since the length of the temporal window can actually be reduced in proportion to the broadening in the spatial dimension. Figure 2 shows an example of such an RAR, clearly revealing both formant structure and individual pitch events.

This is especially useful for the low frequency channels of the auditory model, because the minimum length of the temporal window should really be set inversely proportional to the bandwidth of the respective basilar membrane filter. The low frequency channels have narrow bandwidths, so they would otherwise require excessively long temporal averaging windows. The original RAR often exhibited artifacts in these channels because of inadequate window length (although they were not deemed important because those channels were phonetically uninformative anyway). By way of contrast, the new version has not produced any visible artifacts for any of the data so far analysed.

### 4.3 Frequency Scale

Because the spatial integration reduces the low-frequency resolution of the analysis to 400 Hz, there is little sense in spacing the reference frequencies according to the same ERB-rate scale [6] as was used for the original auditory model. A linear frequency scale is more natural when the resolution is constant, and that is the one which has been used in this paper. This results in a slight over-sampling of the high-frequency region (where the resolution is limited by the bandwidth of the basilar membrane filter), but the recogniser described below is not adversely affected by over-sampling, so this is unimportant.

## 5. RECOGNITION EXPERIMENTS

The task described here is speaker-independent phonetic classification of the phonemes /AE/, /EH/ and /ER/, based solely on isolated frames of RAR data. This task was selected because it is reasonably taxing (these three sounds are often confused by conventional phonetic recognisers and the absence of any context makes the task more difficult still), while the results are easy to interpret because of the small number of classes involved. Furthermore, we would assert that any recogniser which can successfully distinguish these sounds should be able to classify any vowel with a similar degree of reliability.

### 5.1 Data

The data used in these experiments were all the examples of the phonemes labelled /AE/, /EH/ and /ER/ in dialect region 5 of the TIMIT database. A single dialect region was chosen because it was felt that there was no mechanism within the RAR or the recogniser described below which could be expected to cope with inter-dialect variations. A practical recogniser could be expected to incorporate some form of dialect model. Dialect region 5 was selected because it contained a fairly large number of speakers and had the nearest to equal ratio of male to female (albeit still only 37% female).

Both male and female speech have been used because there is no absolute boundary which can be drawn (purely on the grounds of acoustic evidence) between the two. Some speakers exhibit speech patterns which are difficult to classify with any degree of certainty, and if *they* are to be included, how can one justify the exclusion of others merely because they happen to fall more definitely into one or other class? The RAR's spatio-temporal integration is designed to be broad enough to suppress any gender-related pitch variations in any case, while the corresponding formant variations would be better handled by *automatic* clustering of the data (based purely on acoustic evidence) rather than by the use of 'biological' labels, which may group dissimilar acoustic signals together merely because they happen to have been produced by speakers of the same gender.

### 5.2 Processing

Each phoneme was processed with the RAR programme, together with a small amount of the

| intensity | adaptation | frequency | delay | accuracy | ranking |
|---|---|---|---|---|---|
| √ |  |  |  | 56.3% | 8 |
|  | √ |  |  | 51.4% | 11 |
|  |  | √ |  | 57.8% | 5= |
|  |  |  | √ | 58.6% | 4= |
| √ | √ |  |  | 55.6% | 9 |
| √ |  | √ |  | 58.6% | 4= |
| √ |  |  | √ | 56.8% | 7 |
|  | √ | √ |  | 58.6% | 4= |
|  | √ |  | √ | 54.8% | 10 |
|  |  | √ | √ | 57.6% | 6 |
| √ | √ | √ |  | 57.8% | 5= |
| √ | √ |  | √ | 58.6% | 4= |
| √ |  | √ | √ | 59.1% | 3 |
|  | √ | √ | √ | 59.6% | 2 |
| √ | √ | √ | √ | 62.0% | 1 |

Table 1: Recognition results for all combinations of RAR parameters

preceding data (to avoid onset transients). A single frame near the middle of each phoneme was then extracted and used for recognition.

The RAR was calculated using 56 channels in the basilar membrane model, which were then combined to produce 17 output channels, each with a reference frequency separated from those of its neighbours by 200 Hz. The temporal and frequency resolutions were specified as 12.5 ms and 400 Hz, respectively.

### 5.3 Recognition

The recognition was performed by means of linear discriminants [7]. These were calculated to give maximum mean inter-class distance for a fixed (unit) mean intra-class distance, evaluated over the training set [8, pages 40–47]. Each class was then represented by the mean of all the training data for each class, and recognition performed by finding the nearest such mean to each unknown point selected from the test set (in the linear discriminant sub-space).

Some preliminary tests were conducted which showed that recognition performance rarely changed when the number of discriminants was increased above two. This observation was reinforced by observation of the Eigenvalues used to identify the most useful discriminants: only the first two were ever significantly greater than unity, indicating that they corresponded to the only Eigenvectors which produced a useful degree of discrimination. However, the experiments described here used three, just to be on the safe side. Note that this number is lower than would be required for a more general phonetic recogniser, and is only this low because the number of classes is unusually small. It is anticipated that a number closer to eight might be required for more general applications (as was the case in, for example, [7]).

Recognition was performed separately for every possible combination of RAR parameters to assess the degree of redundancy in the data. The results are shown in table 1.

## 6. DISCUSSION

From the results presented here, it appears that a simple linear classifier is probably inadequate for speaker-independent recognition of context-free RAR data frames.

However, the indication offered by the ranking of the various combinations of parameters is quite clear: the more alternative representations are presented to the recogniser, the better. It also appears (from the higher ranked combinations, which seem to follow more definite trends) that it is the amplitude-independent parameters which are most useful (as one would expect), but that absolute amplitude does still significantly improve performance, when presented in addition to the other data.

These conclusions are likely to be applicable to other forms of data as well, and it would be interesting to compare these results with similar ones obtained via non-auditory methods (linear prediction, group delay, filter-bank, etc.).

## References

[1] S. W. Beet and I. R. Gransden. Optimising time and frequency resolution in the reduced auditory representation. In *Proceedings: ESCA Workshop, "Comparing Speech Signal Representations", ISSN 1018-4554*, pages 101–8, ESCA, April 1992.

[2] S. W. Beet, R. K. Moore, and M. J. Tomlinson. Auditory modelling for automatic speech recognition. In *Proceedings: Institute of Acoustics, Speech and Hearing, Vol. 8, Pt. 7*, pages 571–9, IoA, November 1986.

[3] S. W. Beet, H. E. G. Powrie, R. K. Moore, and M. J. Tomlinson. Improved speech recognition using a reduced auditory representation. In *Proceedings: ICASSP-88*, pages 75–8, IEEE, April 1988.

[4] S. W. Beet. Automatic speech recognition using a reduced auditory representation and position-tolerant recognition. *Computer Speech and Language*, 4(1):17–33, January 1990.

[5] M. P. Cooke and M. D. Crawford. The temporal evolution of spectral dominances in an auditory model made explicit. In *Proceedings: ESCA Workshop, "Comparing Speech Signal Representations", ISSN 1018-4554*, pages 131–8, ESCA, April 1992.

[6] B. C. J. Moore and B. R. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, 74(3):750–3, September 1983.

[7] M. J. Hunt and C. Lefèbvre. Speaker dependent and independent speech recognition experiments with an auditory model. In *Proceedings: ICASSP-88*, pages 215–8, IEEE, April 1988.

[8] G. S. Sebestyen. *Decision-making processes in pattern recognition. ACM Monograph Series*, MacMillan, 1962.

# AUTHOR INDEX

ISBN 1 873082 41 X Book 1 (pp1 - 328)
ISBN 1 873082 42 8 Book 2 (pp329 - 651)
ISBN 1 873082 43 6 Set of Two