**Proceedings of the Institute of Acoustics**

# COMBINING AUDITORY REPRESENTATIONS

I. R. Gransden and S. W. Beet

University of Sheffield,
Department of Electronic and Electrical Engineering,
P.O. Box 600, Mappin Street, Sheffield, S1 4DU, UK.

## 1. INTRODUCTION

This paper addresses the problem of combining auditory representations of an acoustic signal, in a manner which preserves the most important aspects of each individual representation. The approach is illustrated with data produced by the reduced auditory representation (RAR) analysis technique [1, 2, 3]. However, it can be generalised to the combination of any spectrogram-like representations. A recognition experiment is described which demonstrates the improvement in noise robustness achieved for an isolated digit database at varying levels of signal to noise ratio (SNR).

Previously, linear discriminant analysis (LDA) [6] has been used for combining different representations of acoustic data. A new method is proposed here, which allows prior knowledge about the importance of different features to be built into a transformation process, simultaneously suppressing pitch information, achieving a higher level of noise robustness, allowing for positional shifts of data within the observation vector, and retaining more information about compact spectral events than is possible with LDA.

## 2. THE RAR REPRESENTATIONS

The RAR is an acoustic signal analysis technique, based on a model of the transformations that occur in the human peripheral auditory system. It is based on the premise that the information coded in the auditory nerve cannot be solely represented by the mean neural firing rate. Other features are assumed to be important: in particular, the temporal information present in the phase-locking of the neuron firings. The phase structure of the basilar membrane displacement, is not apparent from the mean firing rate. The RAR characterises this phase structure by estimating phase derivatives, both with respect to time and cochlear position, giving a detailed picture of local synchrony.

The RAR provides four parameters for each point along the basilar membrane: intensity (related to the signal power in each channel), adaptation factor (a normalised version of the number of neurons expected to be firing at any point), temporal frequency, and spatial frequency of the wave as it passes along the membrane. Each parameter is scaled, on a per-channel basis, based on its expected range of response. All produce spectrogram-like outputs, with peaks corresponding to regions of high energy. These peaks tend to occur in all parameters together, but the correlation is not complete, since some parameters are more

sensitive than others to different aspects of the input signal. All but the intensity parameter are amplitude-independent, and cover a fixed range of values.

## 3. FUZZY SET THEORY

Fuzzy set theory was introduced by Zadeh [8] as means of departing from the hard decision boundaries of traditional set theory. Fuzzy sets indicate a degree of membership in a set, rather than a binary membership decision, which can be advantageous where objects are not distinctly members of one class or another. For a set of observation vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ a fuzzy $c$ partition specifies the degree of membership of each vector in each of the $c$ classes. The degree of membership of $\mathbf{x}_k$ in class $i$ is denoted by $u_{ik} = u_i(\mathbf{x}_k)$ and defined as

$$u_{ik} = \left( \sum_{j=1}^{c} \left( \frac{d_{ki}}{d_{kj}} \right)^{\left(\frac{1}{F-1}\right)} \right)^{-1} \tag{1}$$

where $d_{ki}$ is the Euclidean distance, $\|\mathbf{x}_k - \mathbf{c}_i\|$, between cluster centre $i$ and data vector $k$, and $F \in [1, \infty]$ represents the degree of fuzziness ($F=1$ represents a crisp partition of the data set). The fuzzy membership function exhibits the useful property of bounded membership $\sum_{i=1}^{c} u_{ik} = 1$, where $u_{ik} \in [0, 1]$.

## 4. SELECTING THE DATA SPACE

By appropriate selection of the centres it is possible to characterise the data space fully. In a fuzzy clustering problem the cluster centres, $c_i$, are iteratively updated to characterise the data space efficiently. However, if assumptions can be made about the features which are important in the perception of speech, the centres can be simply positioned to characterise those features, without the computational expense associated with clustering algorithms. For example, if it is reasonable to say that the position of spectral peaks (formant frequencies) can be considered perceptually important, then it would be sensible to position the centres to represent combinations of spectral peaks of the different RAR parameters. All combinations of valleys and peaks in the different parameters must be included amongst the centres so as to provide an alternative membership for the different combinations of parameters. The data space could thus be described by a set of centres for each auditory channel, positioned at the corners of a hyper-cube covering the expected range of each parameter. For a particular spectral event in a particular channel, a large membership function will be associated with the closest centre, indicating significant evidence for that spectral event.

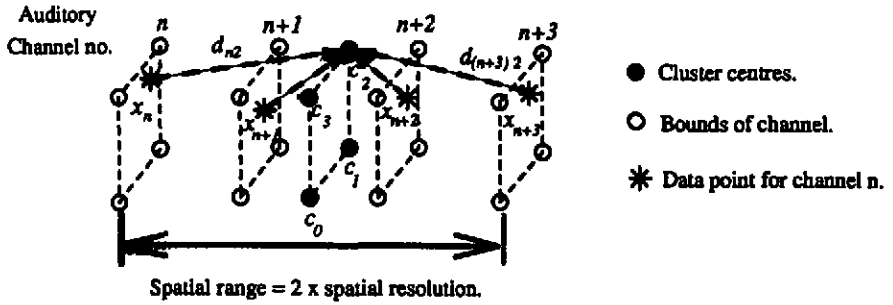


Figure 1: Positioning of cluster centres and definition of spatial range. Each auditory channel is represented by a 2 dimensional auditory observation vector $x_n$.

## 5. POSITIONAL TOLERANCE

It was observed in [2] that the RAR, as with other auditory models, produces representations whose frequency resolution in the low frequency, low bandwidth channels exceeds that which can be effectively used by existing speech recognition algorithms. Intra- and inter-speaker variability can result in vastly different representations of the same utterance. To overcome this problem spatial integration was applied in an attempt to minimise the variation.

Positioning the centres at discrete spatial steps and then calculating the membership function for all data points within a defined spatial range, a level of tolerance to the position of spectral events can be enforced. The spatial position of the centres, either auditorily or linearly based, and the spatial range, affect the way in which the membership functions code the positional tolerance. There is evidence for an auditorily-motivated scale [4] but in this experiment a linear scale has been used [2]. Improved results were obtained by extending the range of the membership function to beyond the required resolution. In this paper twice the normalised range was used. Figure 1 shows a simplified example of the calculation of membership functions for a given spatial range, with each channel being represented by a 2-dimensional vector.

## 6. NOISE FLOOR ADAPTATION

Noise in the acoustic signal will dominate those auditory filterbank channels distant from the major components of the signal. Due to the non-linear nature of the intensity parameter this means that the spectral valleys begin to 'fill in'. This effect can also be seen in the RAR frequency parameters, although for different reasons. For clean speech, channels will be influenced by signal components at or below their centre frequency due to the asymmetric auditory filter passbands. Adding noise will produce a frequency estimate that is related to a complex combination of the filtered signal and noise, dependent upon the relative amplitude of each in the filter passband. This generally results in an increase in the frequency estimates

for the channel. In regions close to the major components of the acoustic signal, channels will be dominated by that component, and should therefore remain relatively consistent. This is true for the SNRs used in this experiment.

The rise in the noise floor can lead to the evidence for spectral valleys being severely weakened, even when they are still obvious in the RAR parameters. It would not be sensible for the centres describing the data space to remain fixed whilst that data space is changing. To overcome this problem, the effective floor of the valleys needs to be tracked. For the experiments performed here, the tracking is performed in a very simplistic way by positioning the centres that form the lower bound of the hyper-cube two standard deviations below the mean of each parameter (across all channels in a particular time frame). Each parameter range is thus independent, with the centres positioned to give the 'best' characterisation of each representation.

## 7. DATA REDUCTION

The information present in the RAR representations is now expressed by a large number of membership functions for the set of cluster centres. The number of membership functions associated with each centre is dependent on the number of channels in the required spatial range. The membership functions explicitly describe the evidence for combinations of spectral peaks and valleys in each auditory filter channel and contain position-tolerant information on those spectral events. However, if this new description is to be of practical use (e.g. interfaced to a recognition algorithm), the data rate needs to be reduced. Redundancy can be expected in the information coded into the membership functions as the presence of evidence for one event will inherently mean the absence of evidence for another event.

For evaluation purposes a simple form of data reduction is used which results in a single spectrogram-like representation. The membership functions associated with a particular centre are averaged to integrate the position-tolerant information spatially. Then, using the eigenvector associated with the largest eigenvalue of the Karhunen-Loéve transform (KLT) [5], a data-dependent optimal linear combination can be achieved with minimal loss of information. Forming the KLT separately for each spatial window, through eigen-decomposition of the covariance matrix of the averaged membership functions for that window, a single value is obtained which is an optimal combination of those membership functions.

The outputs from all the channels are used to form à vector which can be viewed as a spectrogram. Examining the KLT eigenvalues, the percentage mean square error (MSE) introduced by eliminating all but the most significant eigenvector is on average, across all channels, only 12%. A further reduction in data rate can be achieved by applying the KLT across adjacent channels without significant further loss of information, giving, on average, 15% MSE.

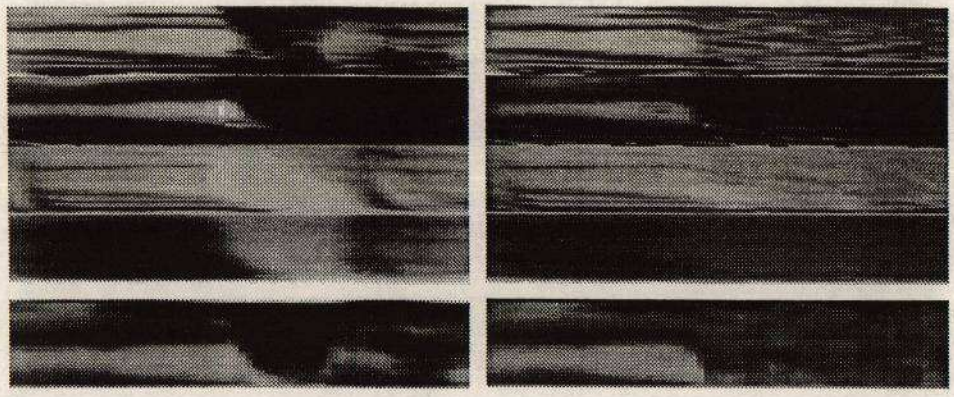

Figure 2: RAR representations (top) (spatial frequency, temporal frequency, adaptation factor and intensity) and combined representation (bottom) for the digit eight, with no additive noise (left) and SNR of 9dB (right).

## 8. RECOGNITION EXPERIMENTS

Due to the amplitude-dependency of the intensity parameter and hence the difficulty in predicting its expected range, it is not included in the fuzzy combination scheme. However, from experiments conducted in [2] it appears that intensity information can be important in discrimination of speech. Therefore the signal energy is calculated (using the computationally-efficient energy operator developed by Kaiser [7]), and is appended to the combined fuzzy membership data.

The three remaining parameters, adaptation factor, temporal frequency and spatial frequency, giving eight cluster centres, are thus combined in the manner described to form a single representation of the auditory data. Figures 2 and 3 show typical examples of the original, separate, RAR parameters and the combined version for the utterances "eight" and "two" respectively. The procedure to generate the single combined representation for each RAR frame is sumarised below:

1. Calculate the RAR parameters.
2. Calculate the position of the lower bound centres for each parameter.
3. Normalise the data space in the Euclidean sense to give all dimensions equal weighting
4. Calculate the membership functions for each centre for all data points within the defined spatial range using (1).
5. Combine information from adjacent spatial windows using the KLT, created by eigen-decomposition of the covariance matrix formed from the training set.
6. Append the Kaiser energy.

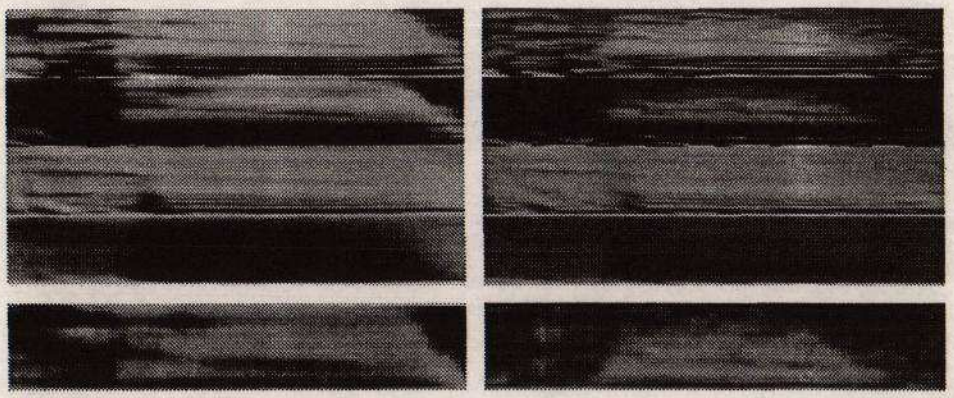

Figure 3: RAR representations (top) (spatial frequency, temporal frequency, adaptation factor and intensity) and combined representation (bottom) for the digit two, with no additive noise (left) and SNR of 9dB (right).

The recognition experiment described here uses a whole word, isolated digit, speech database consisting of five continuous tables, each of 100 digits spoken by a male speaker. Three tables were used for training and the remaining two formed the test set. Pink noise was added to the test data at SNRs of 21dB, 15dB, 9dB, 3dB, and -3dB, and the experiment repeated. Ten-state left-to-right, whole word, multivariate single-mode Gaussian, hidden Markov models (HMMs), with diagonal covariance matrices, were trained on the noise-free data. Single-state noise models were trained to model non-speech sounds (breaths, lip smacks, etc.) and the additive background noise. Using diagonal covariance assumes that all channels are statistically independent. This is not true for the auditory model data, especially in adjacent channels where they are likely to be responding to the same signal components. However, using such an assumption gives a large computational saving for little anticipated degradation in the recognition rates, while allowing more robust estimation of the model parameters.

The RAR was set to give 54 channels covering a frequency range from 50Hz to 5kHz, with a channel spacing approximately equal to 0.5 ERBs. Initial experimentation found that a fuzziness index $F = 2$ and defining the distance measure $d_{ki}$ of (1) to be the $4^{th}$ power of Euclidean distance, produced sufficiently robust results while allowing efficient calculation of the membership functions. Spatially windowing and applying the KLT to the RAR data gave a total of 29 channels plus an appended log energy channel. To allow the RAR to generate sufficient frames to represent the detail of the acoustic data, a frame rate of 5ms was used. The temporal resolution was set to 12.5ms (to suppress pitch modulation) and the spatial resolution was set to 200Hz (the smallest value appropriate to adult male speech). The effective spatial range covers twice that resolution.

| SNR (dB) | RAR | | MFCC | |
|---|---|---|---|---|
| | % correct | % accurate | % correct | % accurate |
| Clean | 100.0 | 97.5 | 100.0 | 100.0 |
| 21 | 100.0 | 98.5 | 98.0 | 98.0 |
| 15 | 95.0 | 93.0 | 78.0 | 78.0 |
| 9 | 88.5 | 83.5 | 28.5 | 28.5 |
| 3 | 43.5 | 36.0 | 0.0 | 0.0 |
| -3 | 0.0 | 0.0 | 0.0 | 0.0 |

Table 1: Comparison of noise robustness of the combined RAR representation with MFCC processed data for decreasing SNR.

The recognition performance of the auditory pre-processor was compared with that of a standard 26 channel mel-scale frequency cepstral coefficient (MFCC) pre-processor with appended normalised log energy. A 25ms temporal window was used with a frame being generated every 10ms (no additional information is yielded by this preprocessor even if a higher frame-rate is used).

Table 1 shows that the combined RAR provide improved representations for the recognition of isolated digits in noise. The RAR performance only begins to fall significantly at SNRs below 9dB, while the MFCC is already severely degraded by this point.

## 9. DISCUSSION

A novel approach to combining auditory representations has been introduced which gives significant improvement in recognition performance for noise-corrupted speech, over that of a traditional MFCC pre-processor. The superior recognition rates obtained using the fuzzy combination of RAR parameters, can partly be attributed to the noise adaptation used. However, the combination method can only provide a consistent representation if the initial representations are robust enough to retain sufficient discriminatory information under noisy conditions.

Most of the errors at moderate SNRs are due to misclassification of the utterance /two/, which can largely be attributed to the weak vowel, which is poorly represented in the original RAR representations, see figure 3. The robustness of the combination is surprising considering the simplicity of the adaptation mechanism which gives only a rough approximation to the noise floor, and ignores temporal continuity in the position of the lower bound centres. A better representation could be achieved by using a noise floor tracking algorithm, and by allowing modification of the upper bound centres to represent weak formants better.

Little evidence of onset or offset information is present in the combined representation. This is not unexpected because, by their nature, these features will only be present for a small fraction of the time and will not be significantly represented in the covariance matrix used to calculate the KLT. The information could, however, have been extracted from the

adaptation parameter, if the matrix had been calculated differently. In particular, an LDA approach could have yielded superior performance if such features were found to be useful in discrimination. The importance of onset/offset information in the discrimination of acoustic data is difficult to quantify, hence the significance of their absence is not known. Calculation of the covariance matrix after variable frame-rate coding, or use of the adaptation parameter, which will be sensitive to onsets, as a weighting function when estimating the matrix, could improve the statistical description of the onsets and hence increase their representation in the output.

## 10. ACKNOWLEDGEMENTS

The authors wish to thank the Speech Research Unit, DRA Malvern, for the CASE award associated with this research and for the use of the digit database.

## REFERENCES

[1] BEET, S. W. : 'Automatic speech recognition using a reduced auditory representation and position-tolerant discrimination', *Computer Speech and Language*, 1990, 4, pp. 17–33.

[2] BEET, S. W. and GRANSDEN, I. R. : 'Interfacing an auditory model to a parametric speech recogniser' *in* 'Proceedings of the Institute of Acoustics, Speech and Hearing', 1992, vol. 14, pp. 321–328.

[3] BEET, S. W. and GRANSDEN, I. R. : 'Time and frequency resolution in the reduced auditory representation' *in* COOKE, M. P. , BEET, S. W. , and CRAWFORD, M. D. (Eds.): 'Visual Representations of Speech Signals', 1993, pp. 175–179.

[4] CHISTOVICH, L. A. and LUBLINSKAYA, V. V. : 'The 'centre of gravity' effect in vowel spectra and critical distance between the formants: psychoacoustical study of the perception of vowel-like stimuli' *in* 'Hearing Research', 1979, vol. 1, pp. 185-195.

[5] FUKUNAGA, K. : 'Introduction to statistical pattern recognition' (Academic Press, 1990).

[6] HUNT, M. J. and LEFEBVRE, C. : 'Speaker dependent and independent recognition experiments with an auditory model' *in* 'ICASSP', 1988, pp. 215–218.

[7] KAISER, J. F. : 'On a simple algorithm to calculate 'energy' of a signal' *in* 'ICASSP', 1990, pp. 381–384.

[8] ZADEH, L. A. : 'Fuzzy sets', *Inf. Control*, 1965, 8, pp. 338–353.