together with a short spectral sample from each transition centre it is possible to obtain a level of recognition which compares favourably with that of two other systems which process the whole signal uniformly.

We are now in a position to consider ways in which the performance of this system could be improved in future. We could, for example:

* make use of further spectral samples from stable spectral regions detected at coherent energy maxima.
* make use of transition cluster patterns as well as spectral samples. These could be modelled in the same way as spectral data.
* increase the amount of training data. Less than 5 examples were found in the training data for over 40% of transitions, and several transitions found in the dictionary were not seen at all in training.
* replace simple dynamic programming with the powerful statistical modelling techniques offered by continuous density HMMs [19]. This could be implemented directly by weighting spectral observations in proportion to their distance from the nearest transition centre. An HMM based system would also avoid the need for hand labelled training data.

REFERENCES


4. A TRANSITION BASED IWR SYSTEM

4.1. Training

We have developed a preliminary transition based IWR system as a test of concept for the idea of recognition based on samples selected automatically from phoneme transitions only. Training consists of detecting transition centres in a corpus of 500 hand-labelled isolated words from each of 5 male and 5 female Spanish speakers. The nearest hand-positioned transition to each detected transition is then found and if this distance is less than 5 frames the detected transition is termed an inter-phone transition and assigned a phoneme-transition label from this transition, e.g. /so/ (see Fig.1a). Otherwise it is termed an intra-phoneme transition and labelled accordingly, e.g. /kk/ (Fig.1b). The expected number of onsets and offsets is estimated in this way for each transition class. Onset and offset occurrences at inter-phoneme transitions are modelled as a Bernoulli process (only 0 or 1 of each may occur, with probability of occurrence = expected number of occurrences for this class, m). For intra-phoneme transitions onsets and offsets are modelled as independent Poisson processes (0 or more of each may occur, with probability of n occurrences = e^{-m}m^{n}/n!). A 15 coefficient spectral sample is taken from each transition centre (see Fig.3). These are used to estimate the parameters for a multivariate Gaussian distribution for each transition class. A note is made of all of the phoneme pairs which are found in the full dictionary.

4.2. Recognition

Dynamic programming is used to calculate the probability of an observed sequence of transition clusters and associated spectral samples having been generated by each of the words in a 2000 word dictionary. An equal weight is given to transition event and spectral probabilities. Transitions not found in the dictionary are given zero probability. Table 3 compares recognition performance with that for two more conventional systems [10] which process the whole signal uniformly. These are DHMM (Discrete Hidden Markov Model) and VQLA. In VQLA each frame is first Vector Quantised using a code-book of prototype MFCC vectors trained for each stable phoneme. This sequence is then smoothed by a majority-filter and passed to a Lexical Access system. Scores are averaged over two speakers (one male, one female) for 500 test words not found in the training set.

Table 3

<table>
<thead>
<tr>
<th>% words correct</th>
<th>DHMM</th>
<th>VQLA</th>
<th>TBIWR</th>
</tr>
</thead>
<tbody>
<tr>
<td>% in top 5</td>
<td>89.5</td>
<td>94.0</td>
<td>95.8</td>
</tr>
</tbody>
</table>

5. DISCUSSION AND FUTURE DIRECTIONS

We have shown that even with a highly simplified model of onset cell function, using just the detected transition sequence
band (see Fig.1) as follows:

- Sample s(t) at 16 kHz.
- Code s(t)→x(f,t) with a 256 point FFT each 10 ms, followed by log compression and MEL scaling into 5 frequency bands.
- Smooth over time, x(f,t)→y(f,t).
- Difference, d(f,t) = y(f,t+1) - y(f,t-1).
- Detect onset/offset at (f,t) if d(f,t) is local maximum/minimum and > 0.6*average |d(f,t)| over time.

2.2. Transition Cluster Detection

Individual transition events are then grouped into separate onset and offset clusters (see Fig.1) using the following rules:

- >= 2 events in cluster.
- <= 8 frames between first and last event in cluster.
- <= 2 frames between adjacent events in cluster.

Parameters controlling detection performance (number of frequency bands, band ranges, degree of smoothing, detection thresholds and clustering criteria) were adjusted to give local optimum detection performance.

2.3. Phoneme Transition Detection Performance

Transition based (TB) detection performance was compared with the performance of the non physiologically based spectral stability measure, previously named as COSH [7,10,17]. The percentage mutual information (PMI) score used here [12] varies from 0 if true and detected transition locations are statistically independent, to 100 if one is completely determined by the other. These scores (Table 1) were not significantly different, and computational cost was also of the same order of magnitude. However, TB detection has the advantage of classifying each transition very reliably as an onset or an offset. The transition cluster pattern can also provide information on broad phoneme identity (see section 3).

Table 1
Transition detection performance for onset/offset cluster detection and the commonly used COSH measure.

<table>
<thead>
<tr>
<th></th>
<th>COSH</th>
<th>TB detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>500 isolated words</td>
<td>19.4</td>
<td>19.5</td>
</tr>
<tr>
<td>50 DARPA phrases</td>
<td>24.7</td>
<td>26.5</td>
</tr>
</tbody>
</table>

3. TRANSITION CLASSIFICATION

Transition clusters were partitioned into those associated only with an onset, those only with an offset and those which sometimes have one or the other or both (offset always before onset). This provided a robust distinction into three very broad classes which was used, together with a spectral sample from the transition centre, for transition classification in the IWR system described in section 4.

It has previously been found [1] that transition cluster shape carries significant information on the identity of the phonemes involved. In order to check to what extent this might be useful for a preliminary broad phonetic classification, we trained a Kohonen map [9] to identify pattern clusters within each of these onset, offset and offset-onset classes. Transition cluster pattern vectors were taken from a 5 band, 8 frame transition...
PHONEME TRANSITION DETECTION AND BROAD CLASSIFICATION USING A SIMPLE MODEL BASED ON THE FUNCTION OF ONSET DETECTOR CELLS FOUND IN THE COCHLEAR NUCLEUS

Andrew C. Morris* & José. M. Pardo
e-mail: a.morris@dcs.shef.ac.uk, pardo@die.upm.es
Grupo de Tecnología del Habla, Dpto. de Ingeniería Electrónica, ETSI Telecomunicación
Ciudad Universitaria s/n, 28040 Madrid
SPAIN

* Andrew Morris is now with the Speech and Hearing Group in the Dept. of Computer Science, University of Sheffield, UK.

ABSTRACT

We present a simple model for onset and offset detection which is based on the broad functionality of onset cells in the cochlear nucleus, the first auditory brain centre. We show that the clusters of transition events detected by this model in the spectrogram can be used to both locate and broad-classify phoneme transitions. A preliminary Isolated Word Recognition system is described which bases recognition solely on evidence from detected transition clusters together with short spectral samples taken from each cluster centre. Recognition performance is compared with that for two other IWR systems of a similar complexity which process the whole signal uniformly.

Keywords: Onset detection, phoneme transition detection, cochlear nucleus, data reduction

1. INTRODUCTION

1.1. The Data-Rate Problem

The data rate from a speech coder which is required to achieve a human level of tone resolution and noise separation is two orders of magnitude greater than that used by conventional speech recognition systems [11(pp.18-20),15]. However, the amount of training data required for valid generalisation theoretically increases in proportion to pattern population size [8], which in turn increases exponentially with pattern dimension. This is possibly one of the reasons why the use of high resolution auditory models for speech coding has generally been found to lead not only to increased computational cost, but also to reduced recognition performance [3,6]. Data redundancy must be exploited to heavily reduce the data volume from an auditory model in some way before it is suitable for direct input to any kind of one-step pattern classifier [2].

1.2. Data Reduction in the Central Auditory System

How might the auditory system cope with this data-rate problem? In the cochlear nucleus (CN), the first auditory brain centre [13,16], the auditory nerve passes in parallel through about seven different neuron types. The firing characteristics for these cells have been established [16]. One of these simply relays the original signal while the others act as various, more or less orthogonal, entropy reducing feature detectors.

1.3. Onset Detection

One of these CN cell types is known as the Onset cell. These have a wide receptive field (about 2 Barks), a high firing threshold and relatively rapid post synaptic potential (PSP) decay. As a result they fire only when they receive simultaneous input across a wide receptive field. After firing they do not fire again until PSP has fallen below a certain “reset” threshold.

In part 2 we present a transition detection model [13, 14, 18] which is loosely based on the functionality of these Onset cells. While there are no offset detector cells in the CN, these can be detected in a very similar way to onsets, and the IWR system described in part 4 makes equal use of both onset and offset detection. Clusters of transition events in frequency and time can then be identified, and we show that these cluster centres are highly correlated with phoneme transition centres. We also show that cluster shapes carry significant information on broad phonetic class [1].

1.4. Subsampling From Diphone Centres

It has been reasoned [4] and shown in practice [5] that short samples taken from phoneme transition centres contain sufficient information for identifying both of the phonemes involved in any phoneme transition. With this in mind, transition detection can be seen as providing a way of focusing the recognition process on concentrations of speech information in the auditory nerve signal, or spectrogram.

1.5. Transition Based IWR

In part 4 we present a preliminary proof-of-concept IWR system which bases recognition on detected transition clusters together with short spectral samples taken from each cluster centre. Recognition performance is then compared with that for two conventional IWR systems.

2. TRANSITION DETECTION

2.1. Onset-Offset Detection

Onset and offset events are detected in each separate frequency