INTRODUCTION

The influence of the phonetic context on the realisation of a word in some continuously spoken phrase may be described by phonological rules. Knowledge of such systematic variation is of potential use in speech recognition systems [1]. However such phonological knowledge is usually expressed as operations on phonetic symbol sequences - whereas speech recognisers operate on continuous acoustic signals. To utilise phonological knowledge previous systems have either used acoustic-phonetic recognisers as front ends [2] or made gross assumptions about the acoustic realisation of phonetic segments [3].

In this paper a procedure is proposed which allows phonological rules to operate on acoustically-derived speech segments. Acoustic segmentation is performed using pair-comparison of test words, and the symbolic labelling of these segments allows construction rules to be written that are similar in form to phonological rules. A combined system of rules may then operate on a phrase grammar to generate a network describing many possible realisations of every phrase but in which the nodes carry acoustic segments instead of phonetic symbols. This network may then be used for recognition with an efficient graph searching strategy such as dynamic programming.

CONNECTED WORD RECOGNITION AS NETWORK TRACING

Connected-word recognition procedures based on whole-word template matching (e.g. [4]) work by determining which concatenation of whole-word templates best explains the input phrase. This process can be viewed as determining which path through a network of templates best explains the input - and this view of the process is useful where the network also captures syntactic (word-order) constraints. Current systems employ a context-free rule to select which word template is associated with each node in the network. That is they embody a model of speech production which supposes that the acoustic realisation of a word is independent of the phonetic context in the phrase. In certain circumstances such an approach works well, not because it is a good model of speech production, but because it is good at differentiating phrases described by the syntax.

CONTEXT-SENSITIVE TEMPLATE SELECTION

A network construction procedure could be designed that chose a template for a node depending on the phonetic context - a context-sensitive rule. Furthermore, context-sensitive rules could be classed as obligatory or optional, hence enriching the transition network with alternative pronunciations. For example, two templates for the word "eight" could be excised from the phrases "eight eight" and "eight two", having the forms [eitʰ] and [eɪt]. A context-sensitive rule would use the first template before words beginning with a vowel and the second template before words starting with a consonant, and might allow either template at the ends of phrases.
Although the context-sensitive selection of word templates could be used to model connected speech effects, the number of different templates required for a word increases as the fidelity of the model increases. What is required is a model of speech production that captures the acoustic regularities and systematic variation in a set of spoken phrases without being confined to whole-word templates. However the selection of suitable templates cannot be performed on an ad hoc basis, because:

1) It is not known what set of acoustic elements is sufficient to describe a speaker’s production of a set of phrases.
2) To use existing phonetic knowledge about systematic variation, the acoustic segments must be given a phonetic interpretation.

**PHONETIC SEGMENTATION**

If it were possible to segment speech automatically into phoneme-length elements which could be labelled with traditional phonetic symbols ([b,e,d,...,etc]) then the context-sensitive template selection rules would be simply phonological rules for connected speech. Rules such as:

- "Alveolar Assimilation"  \[\text{[alv]} + \text{[aplace]} / _ [C,\text{aplace}]\].
- "/r/ Insertion"  \[\emptyset + r / [V,\text{-high}] _ V\].
- "Alveolar Plosive Elision"  \[\text{[alv,plos,avoice]} + / [C,\text{avoice}] _ C\].

E.g. "red bus" as \[\text{[reb\text{\'}bas]}\], "better off" as \[\text{[bet\text{\'}err\text{\'}f]}\], "next week" as \[\text{[nek\text{\'}swik]}\].

A network of phonetic symbols could now be produced from a phrase network by first substituting standard pronunciations for the words, and then enriching the network with optional and obligatory context-sensitive rules of pronunciation variation. The network of phonetic symbols could now be converted to an acoustic template network by the substitution of single templates for each occurrence of a segment. This is precisely how the speech understanding system HARPY [3] organises its linguistic knowledge. The similarity with connected word recognition by whole-word matching is obvious: one determines the best concatenation of word templates to explain a phrase, the other uses a concatenation of phonetic segment templates.

Unfortunately, acoustic-phonetics is not so simple. It is not straightforward to segment and label a speech signal. The acoustic distances between different realisations of one phoneme is not necessarily smaller than the distances between realisations of different phonemes. The transcription of a speech signal as a sequence of phonetic symbols does not seem to be a mechanical process separate from understanding.

**ACOUSTIC SEGMENTATION**

Thus we need an acoustic segmentation process which will provide templates of acoustic regularities in the set of phrases and which allows phonetic labelling of the resulting segments - but without a priori identification of the necessary linguistic units.

The basis for a potentially useful acoustic segmentation technique was described by Moore et al [5]. Two speech signals are time-aligned and the graph representing the distances between the signals is thresholded to separate
acoustically similar stretches. The key to determining acoustic segments that can be phonetically labelled is to compare pairs of words differing by a single phonetic segment (or which only have one segment in common) to excise an instance of a segment and then to compare it with other instances of that segment in different phonetic contexts.

For example, the acoustic segmentation of "pan" might proceed as follows:

- compare "pan" with "pit" to derive segment /p1/ - say [pʰ]
- compare "pan" with "can" to derive segment /p2/ - say [pʰ æ]
- compare /p1/ with /p2/ to derive segment /p3/ - say [a]
- compare "pan" with "pat" to derive segment /n1/ - say [æ ɻ]
- compare "pan" with "din" to derive segment /n2/ - say [n]
- compare /n1/ with /n2/ to derive segment /n3/ - say [n]
- compare "pan" with /p2/ and "pan" with /n1/ to derive segment /a/ - say [æ]

Thus the result of the acoustic segmentation of "pan" might be:

/pl p3 a n3 n2/ - [ph a a i n]

The process can be applied to a large number of test words to derive a multitude of acoustic segments that have been excised from many different phonetic contexts. The number of these segments can be reduced by clustering them in acoustic similarity, and by choosing one exemplar of each cluster. Although it might be expected that segments derived from similar phonetic contexts would cluster together, it is not necessary that exemplars should have a single and unambiguous phonetic interpretation.

ACOUSTIC SEGMENT RECONSTRUCTION RULES

The exemplar segments, being related to portions of the test words with a known similarity, can be used to recreate the test words. The measured error between the original and the reconstruction gives a check on the segmentation process.

If the original test words cover a large enough spread of phonetic contexts, a more ambitious step may now be made: determine which concatenation of segments is the best match for a number of non-test words (words from some task vocabulary, say). The standard connected-word recognition algorithm would be used here. Also determine which segments give the best match to short phrases that exhibit connected speech effects, e.g. "next week" spoken as [nekswik]. Once again, the measured error between original and reconstruction is an indicator of fidelity of matching - the size of the reconstruction error should be less than inter-word distances.

From an analysis of which exemplar sequences describe the phonetic structure of new words, context-sensitive segment selection rules may be constructed that are generalisations of which segment is used in which contexts. For example, if the phonetic form of the word begins [pʰ] then the useful exemplar selection rules might be:

[p] → [p1] / _ [æ]

From an analysis of which sequences match connected speech effects, rules for these effects may also be constructed in a similar form. Thus from an analysis of how these acoustic segments best represent words other than the test words,
THE ANALYSIS AND CLASSIFICATION OF F0 CONTOURS

(F0) and an estimation of the accuracy of the F0 value (CFO). Sections of speech of one second duration were used in both the training and classification.

Variability in the pronunciation of linguistically same stretches of speech can be compensated for by applying a complex, nonlinear time dilation to the speech [8]. The complex, nonlinear time dilation can be decomposed into a series of local time warps; the desired sequence being the one that returns the optimal score in the subsequent cumulative frame-to-frame distance measure applied to the transformed items. The preprocessed speech can be expressed as a sequence of feature vectors

\[ S = s_1, s_2, s_3, \ldots, s_l \]

and the intonation template can be expressed as a similar sequence

\[ T = t_1', t_2', t_3', \ldots, t_J' \]

The distance function \( d \) here uses the self-estimated confidence measure \( c \) in the F0 value \( s \) to form a weighted sum of the squared differences:

\[ d(i,j) = c(i) \cdot (t(j) - s(i))^2 \]

where \( i \) ranges from \( lwb \) to \( upb \) of the data segment
where \( j \) ranges from 1 to \( J \) of the template segment

\( c \) is confidence measure
\( p \) is pattern template F0
\( s \) is speech data F0

The functional equation which equates the policy value for a state at a given stage, is given by

\[ g(i,j-1) + d(i,j) \]

\[ g(i,j) = \min \left( g(i-1,j-1) + 2d(i,j), g(i-1,j) + d(i,j) \right) \]

This is a symmetric form of one of Sakoe and Chiba's DP equations with no slope constraint (that is, \( p = 0 \)). The equation was found by them to perform better than a range of alternatives and has the advantage of being the simplest, computationally. The DP algorithm substantially follows the exposition of Sakoe and Chiba. A modification is used to give a variable geometry adjustment window to constrain the time registration path

\[ |i-(j/s)| < r \]

The initial condition is given by

\[ g(1,1) = 2d(1,1) \]

and the time normalised distance, the optimal score, for a match is given by
RESULTS AND CONCLUSIONS

Considering the results of the two listening groups together (NFP=20), no significant difference between the two types of response categories can be discerned (see fig. 5). The tendency suggested by the pilot experiment is not borne out. The only effect seems to be that the differences between extensions to the 'left' and to the 'right' have been smoothed in some degree. Nevertheless, a striking similarity with the original results in the case of male speaker I (NPP=30) can be noticed (see fig. 2 and fig. 5). In all tests performed in this study the scoring results show the same pattern and the range of mean scores within the nine stimulus conditions remains between about 50 and 70 percent for male speaker I.

If we consider the results of the two listening groups separately (see fig. 6), we find in group one (starting with V responses) in almost all stimulus conditions an increase in correct identification scores in the second test (which included the CVC response category). The other group, however, which performed the two tests in the opposite order, displays an opposite effect: in almost all stimulus conditions the V response option results in higher correct scores than the preceding test with CVC responses. So the tendency to improvement, as indicated in the pilot test, might be the result not of response categories but of the order in which the two tests were presented to listeners.

If in the experiment as a whole the identification scores of the twenty listeners together are grouped according to order of tests (see fig. 7), the second test displays a clear improvement in the identification scores, possibly due to an overall learning effect: the listeners learned in some measure how to cope with the task.

Fig. 7. Percent correct identification in the nine stimulus conditions grouped according to order in which the tests were presented, averaged over the total group of twenty listeners.
THE INFLUENCE OF RESPONSE CATEGORIES ON THE IDENTIFICATION OF VOWELS EXCERPTED FROM CONVERSATIONAL SPEECH

REFERENCES