KEY FEATURES IN CONTINUOUS SPEECH

Mary O'Kane and David Mead

School of Information Sciences and Engineering
Canberra College of Advanced Education

ABSTRACT

'Key Features' is a term coined to indicate blocks of continuous speech which have been recognised with absolute certainty by computationally efficient techniques as having some particular acoustic-phonetic attribute. Key feature recognition is also characterised by no false positives although certain blocks having the features might be missed.

This paper describes the use of the concept of key features as a pivotal element in a project to build a dictation machine accepting continuous speech. A method of locating key features such as voiced speech, voiceless speech, stressed speech, nasals, liquids, plosive bursts, intervocalic voiced plosives and fricatives, voiceless fricatives and the phoneme /s/ is presented and the results of attempting to locate these key features in a passage read by several speakers are given.

INTRODUCTION

In this paper we address the notion that blocks of continuous speech which are recognised with absolute certainty as having some acoustic-phonetic attribute can be used as an integral part of the primary control mechanism of a dictation machine accepting continuous speech input. After describing the role of the key feature concept in an expert's reasoning about an unknown speech passage we go on to describe how reasoning using the key feature concept might be used in building a dictation machine. We then give one method of locating a set of key features and discuss the usefulness of this set of features for the dictation machine design.

A WAVEFORM-READING EXPERT

The aim of the FOPHO continuous speech recognition project [1] is to build a speech recognition system using expertise-capture techniques - in this case the expertise being captured is that of a phonetician transcribing a foreign language. While it is customary to think of the phonetician's expertise as being primarily auditory expertise, experiments by Cole, Rudnicky, Zue and Reddy [2] have demonstrated that certain phoneticians and speech scientists have considerable visual expertise in that they can 'read' spectrograms. In interviewing the expert phoneticians working on the FOPHO project another area of visual expertise was explored - that of 'reading' waveforms [3]. One of the expert phoneticians (P.R.) was particularly good at this and although he generally could not produce a full phonetic transcription of what was said from the waveform alone, he could provide a remarkable amount of information about the phonetic content of the waveform in question. Presented with a section of waveform from an unknown utterance P.R. would first make a series of categorical statements about portions of the waveform which he could immediately identify as having some particular acoustic-phonetic feature. The features identified were sometimes phonemes, a very easy to identify such phoneme being /s/, but very often they were broader phonetic features as as 'voiced', 'nasal' or 'plosive burst'. After making categorical statements P.R. would go on to make a series of more tentative statements about the waveform indicating phonetic features that he believed were probably but not certainly present. However, it is P.R's categorical labellings that are of most interest in this paper. The waveform reading expertise was encapsulated in a set of production rules which were based on a very efficient signal processing technique (described below) which addressed the same waveform primitves that the phonetician used in providing the rationale for his categorical labelling decisions. Initially these rules were merely added to the FOPHO system's knowledge base. However later it was decided that this set of categorical rules might be useful in the primary control mechanism of a dictation system that we proposed building.

A DESIGN FOR A DICTATION MACHINE

The proposed dictation machine, Dicma, is a machine designed to be used for commercial correspondence where a high proportion of the words are predictable. Thus the dictation machine design can make heavy use of a modified word- and phrase-spotting technique. The top-level concept postulated in the Dicma design is that the 'pure' recognition part of the system should produce some form of phonetic encoding (the form of which is discussed below) of the input speech and that this encoding should then be searched for indications the presence of words and phrases which are likely to occur in the dictated material. Sources of such likely-to-occur words are extrapolations from typed-in keywords and studies of the particular user's past correspondence. (For more details of methods used in the prediction of likely words in a dictated passage see [4]). After possible locations of predicted words have been found the presence of these words can be verified or rejected using a test-and-eliminate strategy. A predictive parser can then be used to predict grammatically suitable fillers for the undecoded speech between verified words. In their turn these fillers can be acepted or rejected using the test-and-eliminate strategy.

The basic idea here is not new. The notion of predicting and then searching for likely words in a section of speech was fundamental to the ARPA speech understanding projects [5]. What is being postulated in this paper however is that these ideas can, by a judicious choice of reasoning technique, be pushed a long way for relatively low computational effort and thus enable the production of efficient, low-cost, special-purpose speech recognition devices.

In order to maximise the efficiency of operation of the dictation machine described above we adopted a new structure for the output from the 'pure' speech recognition system component. Generally FOPHO has run according to what is quite a common top-level approach to continuous speech recognition, that is a hierarchial refinement scheme derived from formal phonetic classification theory. An example of this approach is to first classify an unknown sound as either sonorant or non-sonorant, then if it is non-sonorant to see if it is continuant or interrupted and so on. A detailed exposition of this approach has been given by De Mori, Laface and Piccolo [6]. This approach has been generally accompanied by some probabilistic or fuzzy weighting scheme for estimating a degree of belief in any particular classification at any particular level in the hierarchy. However the hierarchical-classification -cum-fuzzy-weighting scheme does not allow us to take advantage of strong categorical inferencing techniques. The issue here is that recognising a particular feature in a stream of speech with near 100 per cent certainty is not nearly as strong a statement as saying that a particular feature has been recognised categorically as being that particular feature. On the basis of this observation we have decided to use two types of reasoning mechanisms in the dictation system - one categorical (or pattern-matching) and the other fuzzy. Categorical reasoning is to be used in likely-word location and a mixture of categorical and fuzzy reasoning is to be used in the test-and-eliminate strategy. This approach is essentially similar to the control mechanism used in many medical expert systems [7].

AN ADEQUATE PHONOLOGICAL ENCODING

To use categorical reasoning for likely-word location the continuous speech input to the system and the set of likely words that are to be searched for must both be encoded according to some robust and adequate phonological encoding scheme. It must be robust in the sense that false encodings must not occur and it must be adequate in the sense that too many ambiguous word locations must not occur. But what constitutes a suitably robust and adequate encoding? First we discuss what might constitute an adequate phonological encoding. Various recent studies on the distributional characteristics of word cohorts that result from encoding complete dictionaries of words according to various phonological encodings (see [8] for an overview) are relevant to this problem. Lai and Attikiouzel [9] carried out a cohort study of Australian English using all the complete (51,018) words of the Macquarie Dictionary [10] as their source. They found that for the various phonological encodings they studied (two of which are close to an encoding we consider below) the expected cohort size is quite small (less than 5) for words of phonetic length seven or greater, certainly small enough to be easily distinguishable with the addition of simple verification techniques. However for words of phonetic length two to six the expected cohort size is rather too high to be easily settled with verification techniques, particularly when it is remembered that the problem we are considering is that of finding words in a continuous stream of speech where false positives can occur across word boundaries.

Accordingly we postulated that a small, but likely vocabulary (such as we would have if we knew the most commonly-occurring words in a user's correspondence) would give rise to a manageable set of words which, when searched for in the input string, should lead to a correct decoding of a fair percentage of the input. To investigate this we carried out a word-frequency study of 33 consecutive letters written by the first author. It was found that 83 words could be classed as high-frequency words [11]. Under quite a weak phonological encoding such as the following:

(voiced), (unvoiced), (vowel),
(nasal), (/s/), (/p/), (/t/), (/k/),

the 83 high-frequency words gave rise to 56 words cohorts, 45 of which were unique. The largest cohort was a five word cohort of phonetic length two. The 56 cohorts were searched for (using rast Bibliographic search techniques [12]) in ten test letters also extracted from the writer's correspondence and also encoded using the eight-class phonological encoding. The high-frequency words accounted for 57% of the words in these letters. A 'best interpretation' of the search on a particular letter was defined to be a reading of the letter that was obtained by assuming that that locations of words of high phonetic length were more likely to be correct than locations of words of low phonetic length. Word locations were not allowed to overlap. With this notion of interpretation, 52% (or 73% if the simplifying assumption of exact phonetic length is made) of the high-frequency words that occurred in the test letters were in the best interpretations of the letters and every high-frequency word was at least in the best or a second-best interpretation. This result together with the smallness of the cohort sizes involved means that even with this weak phonological encoding the verification task to check for correct words and eliminate false positives is quite straightforward. Also as there were no false positives for words of phoentic length six or greater it is probably unnecessary to apply verification to words of this length. Thus for

the environment in which the proposed dictating machine would operate only quite a weak phonological encoding would seem to be adequate to ensure computational efficiency in the operation of the first phase of the machine. But the question that still needs to be addressed is: can such a phonological encoding be achieved robustly for continuous speech?

In the next sections we address this question; first describing the role that key features could play in providing this robust encoding, then giving a means of finding key features and finally discussing the results of running key feature rules over passages of speech read by several speakers.

ABSOLUTE KEY FEATURES

What we call a key feature is a block of continuous speech which has been recognised with absolute certainty as having some particular acoustic-phonetic attribute. In particular, although some sections of speech having this acoustic-phonetic attribute may be missed after the application of key feature labelling rules, one can be certain that there are no false positive labellings, i.e. that no block has been incorrectly labelled. For the categorical reasoning phase of the operation of the proposed dictation machine however we need to find a sub-set of the key features that has the additional property of no false negatives, i.e. all blocks of those features will be found. We will refer to these as absolute key features.

Generally however it should be noted that the indisputable certainty of a key feature label enables immutable anchor points to be established in speech. Around these anchor points a range of hypotheses suggested by phoneme or higher-level prediction rules can be tested using either categorical or fuzzy inferencing. The key feature labels also provide a context which allows context-dependent recognition inferences to be made.

LOCATING KEY FEATURES

When giving explanations for his categorical labellings of printed speech waveforms the phonetician P.R. generally couched them as as arguments about two waveform primitives - zero-crossings per unit time and waveform amplitude. We developed a set of simple waveform analysis techniques which efficiently produces a measure of these two waveform primitives simultaneously. The fundamental procedure of this technique can be described as follows:

> The sum of the absolute values of the valley and peak waveform amplitudes for each adjacent valley-peak pair is calculated as a function (called W1) of the time mid-way between each valley-peak pair.

By repeating this procedure twice on the successive outputs from the procedure we obtain a function (W3) which is similar to the waveform energy [3].

A second procedure is derived from W1. It is referred to as M1 and is obtained by taking the inverse of the time between adjacent W1 points as a function of the mid-point in time between those points. This function can be displayed in a frequency versus time graph.

The third style or procedure used is a function which averages the M1 points for pre-defined window sizes and sampling rates. This function gives (with appropriate choice of sampling rates) a rough guide to various formant trajectories [13]. Graphs of W3 and averaged M1 (for a sampling rate of 20,000Hz) for a female speaker saying "insects may be" can be seen in figure 1. The three procedures described above are all computationally extremely fast.

Several key feature location rules were written which were based on the output from these procedures but which reflected the categorical labellings of the waveform-reading phonetician. This set included rules for the following phonetic labels 'heavy stress', 'voiced', 'voiceless speech', 'nasal', 'liquid', 'voiceless fricative', 'inter-vocalic voiced plosive or fricative', 'plosive burst', 'syllabic peak', 'not high front vowel' and '/s/' and for various broad vowel categories. These rules were written in what is essentially a production rule form. For example the rule for a nasal is the following:

```
label
   name        :  nasal,
   wave        :  speech,
   requires    :  [M1(20000)],
   association :  M1(20000) is long_low_amplitude
                                        (3,300)
end.
```

These rules are written in a special language front-end to Prolog. The declarative style of the rules makes them easy to construct and de-bug. Figure 1 is an example of a typical screen display

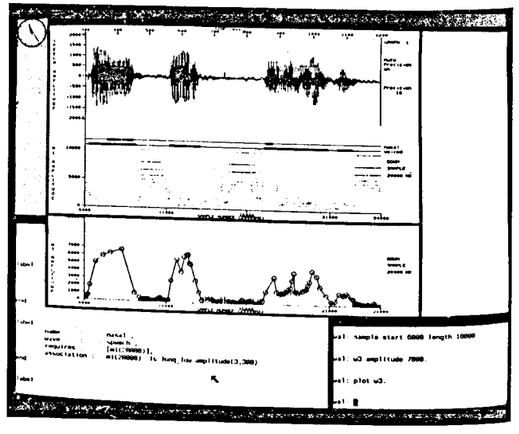

Figure 1: Multi-window screen display for key feature rule system. The top signal processing window displays averaged M1; the bottom signal processing window displays W3. Key features are marked by heavy horizontal lines in the top window.

seen during the running of the rules on a SUN workstation. The central windows are used to graphically display recognition results - the dark lines in the centre show where the key features, 'voiced' and 'nasal' are located - and signal processing results while the window in the right-hand corner is the command window and the window in the left-hand corner is an editor window, the presence of which allows for rapid prototyping of new rules. The various key feature rules written to date were tested on 32 speakers (15 male and 17 female) reading lists of words and on six speakers (three male and three female) reading a short reading passage. Only four absolute key features were found - 'voiced', 'voiceless speech', '/s/' and 'heavy stress'. 92% of the voiceless fricatives and 78% of all the nasals present were located by the appropriate key feature rules. Other key features were located less than 50% of the time using such rules. For key features that are not absolute key features the results varied considerably from speaker to speaker. In particular one of the speakers reading a passage produced plosive bursts less then 10% of the time. However it should be noted that under the condition of stressed speech the key feature 'nasal' is always located 100% of the time.

A ROBUST PHONOLOGICAL ENCODING

From these results it is clear that the current key feature rules would not yield a very strong phonological encoding. The best robust encoding that could be got from the present versions of the key feature rules is:

(voiced), (unvoiced), (stressed vowel), (/s/).

This encoding is even weaker than the eight-class encoding discussed in a previous section. Nevertheless even this encoding does not result in an impossible number of cohorts particularly for words of phonetic length greater than six. Also the fact that at least one key feature becomes an absolute key feature in stressed speech suggests that several encodings should be used when likely words are being searched for. Thus after all key features have been found two phonological encodings of the unknown passage of speech could take place - one the phonological encoding for stressed speech and the other the (weaker) phonological encoding for unstressed speech. The words being searched for could be similarly bi-encoded. First the stressed version could be searched (this would be the faster and more productive search) and then the second, weaker encoding of the unlocated portions could be searched with the weaker encoding of the searched-for words. After that key features which were not explicitly used in the two encodings could be used to eliminate some false word locations. After this still, fuzzy verification strategies could be used.

Furthermore it should be emphasised that key features do not have to be located by the means given in this paper. Any speech segmentaton rule, based on any form of signal processing, that results in no false positives is a key feature rule. Thus a re-examination of speech segmentation studies would doubtless yield a wide range of key feature rules some of which might be absolute key feature rules and this might give rise to stronger phonological encodings for use in the categorical reasoning process.

CONCLUSION

In this paper we have argued that relatively fast and unsophisticated speech processing should yield reasonable recognition rules if categorical reasoning strategies are used.

REFERENCES

[1] M. O'Kane, 'The FOPHO Speech Recognition Project', Proceedings of the Eighth International Joint Conference on Artificial Intelligence, Karlsruhe, 1983, pp.630-632.

[2] R.A. Cole, A.I. Rudnicky, V.W. Zue & D.R. Reddy, 'Speech as patterns on paper', in Perception and production of fluent speech, R. Cole (ed.), Lawrence Erlbaum Associates, Hillsdale, 1980, pp.3-50.

[3] M. O'Kane, J. Gillis, P. Rose & M. Wagner, 'Deciphering speech waveforms', Proceedings of IEEE-IECEJ-ASJ International Conference on Acoustics, Speech & Signal Processing, Tokyo, April 1986, pp.2227-2230.

[4] M. O'Kane, D. Mead, J. Newmarch, K. Byrne & R. Stanton, 'The DICMA Project', Proceedings of the First Australian Conference on Speech Science and Technology, Canberra, November 1986, pp.278-283.

[5] D.H. Klatt, 'Review of the ARPA Speech Understanding Project', J. Acoust. Soc. Amer., 62, 1977, pp.1345-1366.

[6] R. De Mori, P. Laface & E. Piccolo, 'Automatic detection and description of syllabic features in continuous speech', IEEE Trans. Acoustics Speech & Signal Processing, ASSP-24, 1976, pp.365-378.

[7] P. Szovolvits & S. Pauker, 'Categorical and probabilistic reasoning in medical diagnosis', Artifical Intelligence, 11, 1978, pp.115-144.

[8] V.W. Zue, 'The use of speech knowledge in automatic speech recognition', Proceedings IEEE, 73, 1985, pp.1602-1615.

[9] E. Lai & Y. Attikiouzel, 'A comparison of several coarse phonetic classification schemes', Proceedings of the First Australian Conference on Speech Science and Technology, Canberra, November, 1986, pp.361-321.

[10] The Macquarie Dictionary, rev. ed., Macquarie Library, Bee Why, 1985.

[11] M. O'Kane & D. Mead, 'On the feasibility of a continuous speech dictation machine', Technical Report No. 14, Canberra College of Advanced Education, 1987.

[12] A. Aho & M. Corasick, 'Efficient string matching: an aid to bibliographic search', Comm. ACM, 18, 1975, pp.333-340.

[13] M. O'Kane & J. Gillis, 'Efficient derivation of formant-like information from speech waveforms', Proceedings of the First Australian Conference on Speech Science and Technology, Canberra, November 1986, pp.322-327.