A DECLARATIVE APPROACH TO CONTINUOUS SPEECH RECOGNITION

D. Mead and M. O'Kane

School of Information Sciences and Engineering
Canberra College of Advanced Education

ABSTRACT - Arguments for constructing continuous speech recognition systems declaratively rather than procedurally are given, and re-construction of a particular continuous speech recognition system, FOPHO, using a declarative approach is described.

MOTIVATIONS

This paper was initially motivated by the questions "How good does continuous speech recognition have to be?" or "When should we stop building our speech recognition system because it's good enough?" In a sense these questions are easily answered by reference to the task in which the speech recognition system is being used. Thus if a system is being built to recognise strings of digits, then if the system correctly decodes any string of digits given to it one says that the system is a successful recognition system even though the recognition strategy might be little more than sophisticated template discrimination. If one is constructing a spectrogram reading system then one would say one has finished building it only when it can fully extract and decode all the information that an expert spectrogram reader can find. The recognition strategies used in such a system would be much more complex and difficult than those used in the digit recognition example. However an only partially-functioning expert-spectrogram-reader can almost certainly be of use in some speech recognition/understanding applications - a point amply demonstrated in the ARPA Speech Understanding Project (Klatt, 1977). Maybe the questions stated at the begining were ill-posed. A question that reflects our concerns more appropriately is "Even though we might be building a continuous speech recognition system using some particular model of human expertise in this field (spectrogram reader, phonetician or whatever), is it possible to use the partially completed continuous speech recognition system as the recognition part of other speech systems?".

This is a problem that we face regularly in a speech group that is part of an artificial intelligence laboratory which has as its main brief the production of applied artificial intelligence (A.I.) systems. Our primary interest is in investigating techniques for automatic recogniton of continuous speech using models of human expertise as a guide - in our particular case we model a phonetician in the FOPHO system described in O'Kane (1983). When we turn to the problem of constructing applied speech recognition or understanding systems for particular domains, e.g., one current project is to build a dictating machine (O'Kane, Mead, Newmarch, Byrne and Stanton, 1986), we are faced with the decision of whether to construct the new system from scratch or whether to build on top of the FOPHO system. We certainly wish to include the results of the recognition research carried out in the FOPHO project and both types of systems will of course need common signal processing algorithms. However in building applied speech recognition or understanding systems one is generally

trying to do two things:

(a) use domain constraints to turn the recognition task into a verification or pattern matching problem wherever possible,

(b) where (a) cannot be done, use the classical A.I. techniques of heuristic search and automatic inferencing to make deductions from uncertain and noisy information, and make heavy use of rich and diverse domain knowledge.

Perhaps the greatest difficulty using what in many fields of artificial intelligence are seen as standard techniques is the procedural nature of most speech recognition systems (including until recently FOPHO). This phenomenon sets speech systems apart from the majority of other recent artificial intelligence systems which are largely built declaratively. Of course the signal processing basis of speech systems is fundamentally procedural or algorithmic in nature (anyone wishing to be convinced on this point should try to define an FFT declaratively). Although the problems with using A.I. tools in procedural speech systems has been recognised informally for some time, the impetus for moving to declarative structures seems to have come from recent approaches to building automatic speech recognition systems which take the approach of capturing human expertise e.g. the expert spectrogram reader modelled by Zue and Lamel (1986) and by Cabonell, Damestoy, Fohr, Haton and Longchamp (1986) and the expert phonetician modelled by O'Kane (1983).

In reviewing material published on these expertise-capturing systems one often finds that the descriptions of the expertise are given declaratively e.g., Zue and Lamel (1986) give two rules for voicing as follows:

> If the VOT is short,
> and the following vowel is not a schwa,
> then the stop is voiced.
>
> If there is prevoicing during closure,
> then the stop is voiced.

It seems that complex systems motivated by expertise-capture might well be procedural in the signal processing phase but declarative at some later stage.

Thus the declarative approach has a number of advantages when building speech recognition systems and special-domain speech understanding systems:

- it provides a natural way to automate and de-bug the rules that the human expert states he uses when recognising speech;
- it allows the knowledge base of recognition rules to be easily modified and to be built in a modular fashion;
- standard A.I. software tools and techniques can be used;
- recognition systems constructed in this way are easy to integrate with top-down natural language prediction rules such as might be used with a special domain (for a practical example we again refer to the paper on the Dicma project by O'Kane et al. in this volume); and
- the resulting speech understanding systems have a uniform architecture with no enforced structural barriers between the recognition and the syntactic/semantic phases.

INFERENCING MECHANISMS

On the basis of the reasons presented above we decided to re-construct the FOPHO system declaratively apart from the signal processing algorithm phase. However we were influenced in this decision by another factor. This concerned the inferencing mechanisms that we wished to use.

In the past a common top-level approach to continuous speech recognition is a hierarchical refinement scheme derived from formal phonetic classificatory 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 seeing if it is continuant or interrupted and so on. For detailed expositions of this approach see De Mori, Laface and Piccolo (1976) and Weinstein, McCandless, Mondshein and Zue (1975). 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. The initial work on the FOPHO system followed this approach, however recently we have come to the conclusion that the hierarchical-classification-cum-fuzzy-weighting scheme does not allow us to take advantage of strong categorical inferencing techniques. The main issue here is that we believe 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 system - one categorical (we refer to it as 'key feature' recognition) and the other fuzzy.

Thus what we call a key feature is a block of continuous speech which has been recognised with absolute certainty. In particular, although some key features may be missed after the application of labelling rules, one can be certain that there are no false positive labellings, i.e. that no block has been incorrectly labelled.

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 fuzzy inferencing. The key feature labels also provide a context which allows context dependent recognition inferences to be made using fuzzy reasoning.

The separation of the inferencing techniques stresses the correctness of a key feature label in a stronger manner than would a certainty rating of between 95 and 100 per cent. It allows the system to distinguish readily between segment labelling that may be wrong and labelling which can be treated as absolutely correct.

A further advantage to the use of categorical reasoning is that knowledge acquisition can be performed by existing and proven induction techniques. We have yet to explore this possibility in detail.

Key features currently located by the FOPHO system include voiced speech, voiceless speech, silence, stressed vowels, nasal consonants, liquid consonants, plosive bursts, intervocalic plosives, voiceless fricatives, the phoneme /s/ and certain short function words. The performance of the system in terms of not producing false positive labels justifies the

redundancy of some labelling rules which are included in both categorical and fuzzy inferencing schemes. Categorical reasoning and fuzzy reasoning can both be expressed and distinguished very easily in a declarative mode.

DECLARATIVE WRITING OF CONTINUOUS SPEECH RECOGNITION RULES

Analysis of the phonetician's expertise obtained for the FOPHO project demonstrated that the declarative stage of the speech recognition and understanding system should commence immediately after the low level signal processing stage. The signal processing portion of FOPHO consists of a set of algorithms which produce various spectral derivations from the speech waveform. These derivations can be seen as a set of facts on which the recognition system can perform inference.

Two common problem solving strategies need to be applied to the recognition task. A mixture of backward chaining, or goal directed, reasoning and forward chaining, or data driven, reasoning provides a reasonably efficient and semantically powerful framework within which to express the human's recognition expertise. Application of these reasoning strategies to the task of interpreting the spectral derivatives corresponds to looking for particular features (goal directed), while at the same time looking for any feature that the waveform or its spectral derivatives suggest (data driven).

In the current prototype the key feature recognition rules are goal directed. Only a small number of key feature label types are identified, so the use of backward chaining to find these segments is both natural and efficient. Fuzzy rules may be either data driven or goal directed. In this paper we provide details of key feature rules only.

The system requires an overall strategy for the recognition process in addition to locally applied heuristics. Our experience has shown that both the constraining heuristics or meta-knowledge, and the overall strategy can be expressed either declaratively or in a very high level procedural manner. A major benefit of this approach is the flexibility of the system, with the ability to change global strategy without resorting to writing C code or Unix shell scripts.

The approach outlined here is useful in the evolutionary stage of a continuous speech recognition/understanding system. Although efficiency is a primary concern, and indeed at the signal processing stage the algorithms used are extremely fast, we do not propose that this form of implementation would work in real time. Rather, conversion to a real-time system will be done in the future.

The rules for recognising key features are descriptions of the distinguishing characteristics of those features. They are categorical although possibly non-monotonic rules that describe time-position-relationships between attributes of the spectral derivatives. They also perform a bridging function between the algorithmic processing stage and later declarative stages, and often refer to some procedural knowledge. An example of a descriptive key feature rule is:

> Where a portion of the speech waveform is labelled as voiced but the energy shows a sudden dip, that portion is labelled "inter-vocalic-voiced-plosive".

The definition of a sudden dip may well be best expressed procedurally.

In building a program to perform inference on the speech data we had the option of creating a special purpose language using obvious non-terminal symbols derived from hierarchical classification schemes referred to above. An interpreter and inference machine for this language could then be built using standard software tools and a fast procedural language. We quickly discarded this option because it would produce an inflexible, almost hard-wired system which could prove to be inappropriate after a short period of use.

We chose instead to use the general purpose declarative language Prolog. This language provides a suitable environment for rapid prototyping of solutions to problems which involve categorical first order logic reasoning. It can be extended easily to fuzzy and probabilistic reasoning schemes. Programs in Prolog can also be procedural and so the link between the algorithmic signal processing and the descriptive feature recognition can be established smoothly. A disadvantage of this approach is the slow speed of the system. However this is not critical at this stage of development.

At present, development is proceeding in a Unix environment. The multi-tasking capability and the ease of interprocess communication facilitate the movement of data between the signal processing stage and the recognition stage. It also allows us to approximate the behaviour of an eventual real-time system during this development stage.

The continuous input stream of facts from the signal processor is divided into time windows for inferencing purposes. The window size and window overlap is chosen to match the time constraints stated in the recognition rules. This choice can be predetermined or derived automatically by the rule interpreter. At present, a window size of about one second with overlap of about 250 msec is used. With correct choice of window size and a suitable scheme for dealing with conflicting decisions made near window boundaries, the scheme enables the decomposition of an infinite task of continuous recognition into finite chunks more amenable to processing by humans and by computing machinery.

The window approach also permits a simple representation of data within the inference engine. Input data is represented as separate lists of tuples for each spectral derivative. A tuple typically consists of a pair of x-y coordinates for a point on a waveform. Output data consists of further lists of tuples for the window. A separate list for each label type is produced, with a member containing a vector of attributes describing the labelled portion, including the starting and finishing points and possibly a justification for the labelling decision.

The task of the inferencing program is to find and label all segments of the speech waveform within the current window which match the descriptions contained in the key feature rules.

Top level rules are written declaratively, for example

```
where voiced
    contains sudden_dip_energy
label inter_vocalic_voiced_plosive.
```

These rules often refer to attributes which are best defined more procedurally, for example, sudden_dip_energy. In these cases, the attribute is defined using Prolog code to describe a single instance of that attribute. The rule processor accepts these Prolog definitions as well as the rules written in the above format and expands them to an internal format which allows all instances of the segments to be labelled within a window.

SUMMARY

With regard to our concern about how good a continuous system has to be, we do not believe that we have answered this question but we do believe that we have adopted a system architecture that will allow us to investigate the effectiveness of a large and varied number of continuous speech recognition and inferencing rules in a modular, and hence easily extendable, fashion. The declarative architecture also allows us to use our recognition system within special-domain speech understanding systems even while the recognition system is still under development.

REFERENCES

CARBONELL, N., DAMESTOY, J., FOHR, D., HATON, J. & LONCHAMP, F. (1986) "Aphodex, design and implementation of an acoustic-phonetic decoding Expert System". Proceedings of IEEE Int. Conf. on Acoustics, Speech & Signal Processing, Tokyo, 1201-1204.

DE MORI, R., LAFACE, P. & PICCOLO, E. (1976) "Automatic detection and description of syllabic features in continuous speech". IEEE Trans. Acoustics Speech & Signal Processing, ASSP-24, 365-378.

KLATT, D. "Review of ARPA Speech Understanding Project", J. Acoust. Soc. Am., 62, 1345-1366

O'KANE, M. (1983) "The FOPHO Speech Recognition Project". Proceedings of the Eighth International Joint Conference on Artificial Intelligence, Karlsruhe, 630-632.

O'KANE, M., MEAD, D., NEWMARCH, J., BYRNE, R. & STANTON, R. (1986) "The Dicma Project". Proceedings of the First Australian Conference on Speech Science & Technology, Canberra.

WEINSTEIN, C., McCANDLESS, S., MONDSHEIN, L. & ZUE, V. (1975) "A system for acoustic-phonetic analysis of continuous speech". IEEE Trans. Acoustics Speech & Signal Processing, ASSP-23, 54-72.

ZUE, V. & LAMEL, L. (1986) "An Expert Spectrogram Reader : A knowledge-based approach to speech recognition". Proceedings of IEEE Int. Conf. on Acoustics, Speech & Signal Processing, Tokyo, 1197-1200.