THE FOPHO SPEECH RECOGNITION PROJECT

Mary O'Kane

School of Information Sciences
Canberra College of Advanced Education
P.O. Box 1, Belconnen, 2616. Australia

ABSTRACT

The FOPHO (F_oreign Phonetician) speech recognition project concerns the development of a system to produce a reasonably high quality phonetic transcription output from continuous speech input. The system is developed to perform in a way which approximates the actions of a phonetician trying to transcribe a foreign tongue, (in the case of FOPHO, Australian English). Because of this central philosophy, FOPHO is a very interactive system and has facilities for automatic learning and analysis of its own performance. Good quality recognition is achieved through algorithms which are very context-dependent and which are sensitive to a variety of possible productions of similar sounds even though the system itself is speaker independent.

Introduction

The aim of the FOPHO project is to build an automatic system which can learn to recognise continuous, conversational-style speech, unrestricted by speaker or context, in the manner in which a Foreign Phonetician (hence the name FOPHO) would learn to understand a language. The novelty of this approach lies in the system design which combines techniques developed in the Artificial Intelligence fields of Learning Systems and Expert Systems with the principles of Automatic Speech Recognition. The result is a system which is a 'Learning Expert' and that is precisely what a phonetician trained to approach an unknown language with a view to identifying and categorising its sounds is.

In practice FOPHO is designed to recognise Australian English which is a dialect of English more suited to automatic recognition than many other dialects of English because although it is spoken by a large number of speakers (approximately fourteen million) it is remarkably free of regional variation. (Mitchell and Delbridge, 1965).

This paper describes work done partially in the Department of Engineering Physics, Research School of Physical Sciences, Australian National University. It was supported by grants from the Australian Computer Research Board.

System Features

In this section we briefly describe the main features of the FOPHO system apart from the actual speech recognition strategy which is described in the next section. It must be emphasised that FOPHO is envisaged as a long-term project with the system providing a framework for its own development. Thus, while is is primarily meant as an automatic speech recognition system it is also designed to provide an automated framework for experiments in phonetics and for experiments in automatic-speaker characterisation.

The expertise of FOPHO resides in its recognition algorithms which are designed to be as detailed as possible. i.e. context-dependent rules and rules dealing with a wide range of (often redundant) phonetic features are included in the algorithm wherever possible. Each recognition algorithm consists of conjunctive and disjunctive concatenations of simple fuzzy breakpoint rules (Demichaelis, De Mori, Laface and O'Kane, 1983). Each algorithm slots into a hierarchy of recognition algorithms (segmentation algorithms precede fine classification algorithms and so on). This hierarchy provides a chaining between all the various recognition rules and through this chaining the system can carry out an analysis of its own performance and, when requested, modify its own recognition rules. The system provides the analysis of its own performance in the manner and terms of a phonetician.

With these considerations in mind the main features of FOPHO are:

1. the system is interactive, i.e. it can answer back using its synthetic voice which is the Klatt cascade/parallel formant synthesizer (Klatt, 1980), and it has simple yes/no question-answer routines for gaining general information about a new topic or a new speaker. These routines can be run through keyboard entry or by speech response using simple template-matching speech recognition techniques. FOPHO also has been provided with the ability to produce IPA phonetic script on a graphics screen (Millar and Oasa, 1981) and, in its phonetics experiment mode, it can accept graphical input through a graphics tablet.

Collectively these interactive routines are known as the 'friendly face[1] of FOPHO;

2. the system seeks verification, i.e. using in-built facilities for synthetic speech and phonetic symbol output the system seeks verification of what it 'thought it heard[1];

3. the system can be made self-tuning. In this learning mode it can be made aware that it has made a recognition mistake and it will seek to identify and then modify the algorithm which has caused it to make a mistake;

4. the system is modular in design, i.e. the algorithms of the system are written as discrete modules so that they can be individually and independently tested and modified and so that the system can perform at least partial recognition even though a complete set of recognition algorithms has not been developed. Thus, as the system stands at present it can segment speech into broad phonetic classes, it can perform reasonable classification of vowels, and it has very robust algorithms for the fine classification of plosive consonants and liquid consonants. However, its fine classification algorithms for other consonant classes are still under development. At present if it is asked to recognise a phrase, for example:

   Fred sings

   it will produce as its highest-weighted fuzzy output something like the following:

   (fricative)(r) (mid-high front vowel)(d) (fricative)(high front vowel) (nasa1)(fricative);

5. the system has on-going speaker adaptation, i.e. when a new speaker is introduced to the system, the system creates a file for that speaker in which it deposits on-line-gathered statistics about that speaker's pronunciation for later use in conjunction with the general speaker-independent recognition algorithms;

6. the system, which at the moment is being developed as a speech recognition system, has the facility for being extended to act as Speech Understanding System according to the same general model except that the phonetician is replaced by a linguist.

The principles on which the model is based should lead eventually to reasonable recognition of unrestricted continuous speech; these principles can be also used in the design of special-purpose speech recognition systems.

Principles of Recognition

Two basic principles underlie the speech recognition rules used in FOPHO. They are:

1. context-dependent recognition rules are better than context-independent rules;

2. perfect recognition is impossible to achieve so that rather than produce one classification for each phoneme it is generally better to produce a list of possible classifications with associated fuzzy weightings.

The recognition process starts when the digitized, LPC analysed speech is segmented using a modified version of the hierarchical segmentation algorithm developed by De Mori, Laface and Piccolo (1976). After this, primary vowel recognition takes place. This is done by a context-independent algorithm which relies mainly on formant and timing data to achieve results. As there are twenty vocalic nuclei in Australian English, it is quite common that a vowel can receive a classification with a high fuzzy rating in several categories at once. For example, a vowel that is actually / I / might receive classifications of over .6 (on a scale of 0.0 to 1.0) in the categories / i /, / I / and /r/. Because of this and because fine vowel classification is not needed for the context-dependent consonant recognition algorithms (this primary vowel recognition providing the 'context' for the consonant recognition algorithms) the vocalic sections are also given ratings in broader categories such as high front vowel, low back vowel etc.. Of course if some data for the current speaker is already held knowingly by the system in its previously-met-speaker files these can be consulted and finer vowel recognition achieved. It should be noted also that the vowel recognition algorithm is more successful for female voices than for male voices because of the wider spread of the second, and third formants in female voices. Further details of this and other features of this vowel recognition algorithm are given in O'Kane (1981).

After vowel recognition is completed, fine classification of the consonantal parts of the utterance is undertaken. It has been established that the consonants in conversational Australian English show strong coarticulatory effects due to both anticipatory and carryover articulation, and that the presence of a word boundary does not significantly inhibit either type of coarticulation unless there is a long time gap between the words (O'Kane, 1981). Thus, we have found that consonant recognition is most successfully done using context-dependent algorithms, and not surprisingly this type of recognition is most successful when a consonant occurs in a VCV context. However, even when only VC or CV coarticulatory rules can be used, due to the presence of consonant clusters, the results are very satisfactory.

The problem with consonants in rapid conversational Australian English is that they can be produced using several features or using only a subset of those features. For example, a

perfectly produced plosive consonant will have well-defined formant transitions from and to surrounding vowels as well as a burst of a certain spectral shape, and the Voice Onset Time will be of a certain length all depending on the place of articulation of the consonant and the nature of the surrounding vowels. But certain speakers at certain times use ony some of these cues, for example, never producing a /d/ burst before a high back vowel or having negligible formant transitions before /b/. As the algorithms for recognising plosives are designed to be speaker independent these effects have to be allowed for. Also it has been found necessary to have different consonant recognition algorithms for male and female voices - at least algorithms that differ when dealing with frequency dependent features. How an algorithm is developed is described briefly in the next section.

After fine consonant recognition has been completed for an utterance the results for each phoneme are reviewed. If a phoneme has received no fuzzy rating in any category greater than 0.5 or if it has a high rating in several categories simultaneously the possibility of a mistake at lower levels is considered, and a certain amount of backtracking is performed. The system displays its results to date, and an analysis of these results, indicating particularly those parts where no successful classification has been achieved and where rule modification is perhaps necessary.

After this the results can be further improved by operating the (phonetic) context-dependent and prosody-dependent word boundary algorithms which allow some word and phrase boundaries to be tentatively tagged.

Design of Recognition Algorithms

The design of the speech recognition algorithms is accomplished by first considering what phonetic features are known to be perceptually important in the human recognition, and which parameters have been shown to be particularly robust in the automatic recognition of a particular class of sounds. This information is used as a guide to the basic form of the algorithm. Then a large number of examples of all members of the sound class being recognised in all possible contexts, and as spoken by a large number of sneakers is considered, and from this information the final recognition rules are developed. As F0PH0 is designed to recognise continuous speech, the sound examples considered in developing the rules are extracted from continuous speech. In order to obtain these sound examples in all possible contexts a special experimental paradigm has been devised whereby speakers are presented with lists of words or word pairs containing some target sound in all possible contexts, and are asked to put these words or word pairs into sentences as rapidly as possible. This has proved to be very useful in obtaining the required material in a continuous speech format.

To provide a wide range of material to further develop and test the FOPHO algorithms a large database of spoken Australian English has recently been collected. Thirty-two volunteers (fifteen male and seventeen female) have provided, on ten different occasions, exemplars of a variety of speaking styles ranging from the formal (set reading passages, reading lists of sentences and lists of words) to the informal (word games of the type described above, free conversation). This material is also intended for use in speaker characterisation research (O'Kane, Millar and Bryant, 1982).

Conclusion

The FOPHO speech recognition project which aims at achieving high quality recognition of conversational Australian English has been described. This project is expected to take some years to complete and has thus been designed in a very flexible way so that the project can take advantage of new developments in Artificial Intelligence.

References

P. Demichaelis, R. De Mori, P. Laface, and M. O'Kane, 'Computer recognition of plosive consonants using contextual information', IEEE Transactions on Acoustics, Speech and Signal Processing, 1983, In press.

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

D.H. Klatt, 'Software for a Cascade/Parallel Formant Synthesizer', Journal of the Acoustical Society of America, 67, pp. 971-995, 1980.

J.B. Millar and H. Oasa, 'Proposal for ASCII coded phonetic script', J. Int. Phonetic Assoc, _2, pp. 62-74, 1981.

A.G. Mitchell and A. Delbridge, 'The Speech of Australian Adolescents: A Survey', Angus and Robertson, Sydney, 1965.

M. O'Kane, 'Acoustic-Phonetic Processing for Continuous Speech Recognition', PhD Thesis, ANU, Canberra, 1981.

M. O'Kane, J.B. Millar and P. Bryant, 'Design and Collection of an Australian English Database Preliminary Report', Technical Note No. 5, School of Information Sciences, CCAE, Canberra, 1982.