# THE DICMA PROJECT

M. O'Kane(*), D. Mead(*), J. Newmarch(*), R. Byrne(**)and R. Stanton (***)

(*)School of Information Sciences and Engineering
Canberra College of Advanced Education

(**)School of Liberal Studies
Canberra College of Advanced Educcation

(***)Department of Computer Science, The Faculties
Australian National University

ABSTRACT - The outline of a speech understanding project which has recently commenced is presented. This project involves the construction of an automatic dictating machine which will accept continuous speech input. A notable feature of the system is its reliance on information it generates from user-provided keywords.

## INTRODUCTION

The aims of the Dicma project are twofold:

1) to build a useful office automation system which uses continuous speech recognition technology,
2) to explore the general principle of making subject- and speaker-independent continuous speech recognition technology viable by provision of adequate top-down information.

When completed the Dicma system will process a letter dictated to it and then will produce the letter on a printer or file it for later use with a conventional word processor. Dicma is a continuous speech recognition system which is provided with top-down help generated initially from some key words that the user types in. To see how the system might work consider the following dialogue which could take place in the office of some mythical company called, let's say, Dicma Office Automation Systems:

SYSTEM (speaking): Hi, I'm Dicma your new spoken-word processor. If you dictate a letter I'll transcribe it and produce a printed copy ready for you to sign and send. I'll need you to type in a couple of things, but most of our communication will be by voice. Ready? If so please press the start button on the handpiece.
Thank you. Please type in the name of the person you want to write to, then the postcode or the suburb.
USER (types: "Mr Bus Nissman... 2601")
SYSTEM (speaking): Please tell me what suburb Mr Nissman lives in.
USER (speaking): O'Connor.
SYSTEM (speaking): Please tell me his street address.
USER (speaking): 45 Wattle Street.
SYSTEM (speaking): Shall I say "Dear Mr Nissman" or "Dear Bus"?
USER (speaking): Dear Mr Nissman.
SYSTEM (speaking): Please type in some keywords that you want to use in this letter.
USER (types) "enquiry", "Dicma", "folder".
SYSTEM (speaking): Thank you. Now please begin dictating. If you want me to start a new paragraph, push the "para" button on the handpiece.
At the end, press the "end of text" button. Waiting...

USER (speaking): Thank you for your letter enquiring about office automation systems. Our Dicma system, which is easy to use and competitively priced, would be ideal for your requirements. Full details of this innovative product can be found in the accompanying folder.
SYSTEM (speaking): Will I use your normal sign-off?
USER (speaking): Yes
SYSTEM (speaking): I've put your letter on the screen. I have highlighted one passage I can't manage. Could you look at it and fix it up on the keyboard please?
USER (types in whatever corrections are needed)
SYSTEM: Thank you. Your letter is now being printed.

For entering material such as address details, single word and limited vocabulary word discrimination techniques are indicated. The operation of the techniques used in transcribing the body of the letter are more complex. A possible mode of operation is the following.

GENERATING POSSIBLE WORDS

After the keywords are entered the system would build up from them a large list of words that might appear in the letter. Consider the keyword "enquiry" in the example above.

From a dictionary (Concise Oxford) the following entry is found

**enquire, enquiry.** See INQUIRE, INQUIRY.

**inquir'ỹ**, en-, (in-kw-), n. Asking; question; investigation; *make ~ies*, ask (*about* etc.); *court of ~y* (investigating circumstances of mishap etc.). [f. prec. +-Y⁴]

Thus we have alternative words: "asking", "question", "investigation" and some phrases: "make enquiries", "ask about", "court of enquiry".

From a Thesaurus (Roget) we get a much larger set of words and phrases some of which are, of course, highly unlikely to occur in a business letter.

**461. Inquiry.** [Subject of Inquiry. Question.]—N. inquiry; request etc. 765; search, research, quest; pursuit etc. 622.

**765. Request.**—N. requ-est, -sition; claim etc. (*demand*) 741; petition, suit, prayer; begging letter, round-robin.
motion, overture, application, canvass, address, appeal, apostrophe; imprecation; rogation; proposal, proposition.
orison etc. (*worship*) 990; incantation etc. (*spell*) 993.
mendicancy; asking, panhandling, begging etc. v.; postulation, solicitation, invitation, entreaty, importunity, supplication, instance, impetration, imploration, obsecration, obtestation, invocation, interpellation.

Of course one can also use other dictionaries and specialist dictionaries such as crossword dictionaries to learn more possible words and phrases that might occur as alternatives to or in conjunction with some particular keyword.

Another source of information about possible words are rules giving word variations. Thus we might consider the plural of enquiry - enquiries, the verb deriving from it - enquire, and parts of this verb - enquires, enquiring, enquired.

Yet another source of information is a knowledge of letter structure and protocol. Thus the first keyword and its variations are likely to occur in combination with common starting phrases of letters which include:

| A | B | C |
|---|---|---|
| Thank you for | your letter | regarding |
| Thank you for | your note | of the [date][month] |
| With reference to | your letter | regarding |
| With reference to | your phone call | of the [date][month] |
| I refer to | your request | |

As an aside here two things should be noted

1) that the start of a letter depends very much on the intention of the sender of the letter. Thus maybe Dicma should ask the user whether the letter is to be rude or polite. Is it to be a letter answering a query or is it a letter requesting/demanding information?

2) The way a letter begins is very idiosyncratic to the sender. Thus when Dicma is being introduced to a new office, a Dicma sub-system could be used to analyse automatically correspondence that has been sent out over a period of time. It would learn what are the common words and phrases used in the office correspondence and what are common letter beginnings and endings.

On the issue of keywords generally it should be noted that some keywords, such as proper nouns, would give rise to an extended sets of words and phrases. Consider the keyword "Dicma". This is a proper noun and this keyword would be known to the system from a local knowledge source which would store with the word "Dicma" words and phrases describing its attributes e.g.

| Proper Noun | Attributes |
|---|---|
| Dicma | Easy to use, Competitively priced, Is a desk-top system, Easily adapted to a new application, Price is $A5675, Comes in a range of colours |

The way the system would learn about this proper noun and its attributes is via another Dicma sub-system which could be run whenever proper nouns relevant to some particular application are to be added to the system. Considering the Dicma Office Automation Systems example, as well as giving the system the attributes of Dicma it would be told the name of managing director and his attributes e.g. "live-wire", "go-ahead".

With some keywords, still other sources of information might be important. One of these is a knowledge of general and specific categories. Thus Dicma is a specific example of the (more general) class of office automation systems; folders are a specific example of the (more general) class of printed matter.

Now let us consider what happens when all the words and phrases evoked from the keywords and from general-purpose knowledge sources such as letter protocol sources have been obtained.

PHONETIC PREDICTION

The next step the system would take is to use the set of words and phrases it has generated to predict phoneme patterns that might occur in the letter being dictated. Obviously the first thing to do is to convert all the words in the generated set into phonetic form. This might be done, for example, using the Macquarie dictionary which provides phonetic transcripts of words in Australian English:

**automation** /ɔtə'meɪʃən/, *n.* 1. the science of applying automatic control to industrial processes; the replacement of manpower by sophisticated machinery. 2. the process or act of automating a mechanical process. 3. the degree to which a mechanical process is automatically controlled. [b. AUTOM(ATIC) + (OPER)ATION]

It is important to note however that words in continuous speech are not always produced as they are in citation form (the form presented in the dictionary) for example the phrase:

Can you understand it?

might, in continuous speech, be said as

kənjuʌnəstænət

with neither of the d's in "understand" being produced. Thus for every phonetic transcript of every word in the set of words generated from the keywords, alternatives would be produced which allow for the effects of continuous speech such as vowel reduction (particularly in unstressed words) and omission of certain consonants. Many of the rules for generating these alternatives come from research on continuous speech recognition. Other rules would come from text-to-speech rules such as those used in text-to-speech systems with synthetic speech output. Having generated lists of alternative phonetic possibilities for all the words and phrases in the generated set, higher-level versions of all these phonetic representations should be generated to facilitate early matching with initially processed input speech. Thus for example the word, "Fred", which in citation form is produced as

fred

would have the higher-level representation

(voiceless fricative)(liquid)(vowel)(voiced plosive)

and the still higher representation

(voiceless section)(voiced section)(possible burst)

These higher-level representations are easily generated from a knowledge of phonetic categories. Also generated at the time of generating these higher-level representations would be maximum and minimum estimates of the duration of each phoneme and voiced or voiceless section.

Now we turn to the question of how to use this information to match the incoming speech.

ANALYSIS AND MATCHING OF INCOMING SPEECH

As the user dictates the letter the speech will be digitised. This

digitised speech will then be stored (for future reference) and analysed using a variety of techniques.

Following the approach taken in the FOPHO project (O'Kane, 1983) modified by recent work (Mead and O'Kane 1986), the segmentation rules developed by O'Kane, Gillis, Rose and Wagner (1986) are used to locate the portions of voiced and voiceless speech and silence. These rules also locate all stressed and many unstressed vowels and /s/. After this most nasal sounds, many liquids, many intervocalic plosives and fricatives and all plosive bursts are located. As well possible function words of the type (voiceless sound)(vowel)(voiceless sound) are tagged and vowels are recognised in broad categories (high front, low back etc.). At this point a lot is known about the phonetic context of the message although of course there is a lot more that can be found out.

Thus this is a good point to commence the matching process or, more correctly, the word spotting process. This process involves the matching of high-level phonetic descriptions with their associated timing estimates from the generated word set with the analysed input speech. This can be done quite quickly using an algorithm such as the bibliographic search algorithm developed by Aho & Corasick (1975). Where a potential word match has been found a test-and-eliminate strategy can be undertaken at the phonetic level to confirm or deny the match. Even if a match is denied through this process, a lot is learnt about the possible identity of the portion of speech with which the match was attempted. For example if a voiceless sound has been proven not to be a plosive then the system can conclude that the sound must be a fricative. Phonetic deduction rules of this type would be included.

SYNTACTIC ANALYSIS AND PREDICTION

As more and more words and phrases are confirmed as definitely occurring in the input speech, attempts would be made by the system to recognise on one hand, and predict on the other, the words occurring in between the recognised 'islands'. This could be done using predictive parsers based on either definite clause grammars or augmented transition-network grammars. The predictive parser could be used to predict the syntactic constructs that would be appropriate between recognised words. Attempts would then be made to find words belonging to the appropriate syntactic categories. This search would be helped by constraints imposed by the signal processing/phonetic analysis and by the elimination strategy and the phonetic deduction rules outlined briefly above. Thus, for example, if an appropriate syntactic filler between two recognised islands is a noun and if we had the constraints that it is of the form

(voiceless fricative)(vowel)(nasal)

then possible words would be "fan", "ham", "shin".

These words could then be put through the test-and-eliminate strategy mentioned earlier.

SEMANTIC CHECKING

Another possible addition to the system would be a set of rules to perform simple semantic checking in order to see if the identified valid syntactic structures "made sense". In this respect a fairly simplistic semantic

checker, say an extension of the Quillian Semantic Net (1968), would probably be adequate at least in an initial version of the system.

SYSTEM ASKING FOR ASSISTANCE

When the system had to give up on trying to recognise a passage it could ask for help either by keyboard interaction as indicated in the example at the beginning or, in a more sophisticated version of the system, by voice. There seems nothing wrong with this notion as long as the user is not asked to supply too much help. After all, even the best steno-secretaries have to ask occasionally for phrases to be repeated when they are taking shorthand. Admittedly the situation is slightly different here as the request for assistance from the system would be made after the complete letter is dictated.

FEASIBILITY OF BUILDING THE DICMA SYSTEM

None of the parts of the Dicma System could be said to be revolutionary in Artificial Intelligence terms although some of the components of the proposed system rely on leading-edge research in the field. The novelty of the proposal lies in the bringing together of a range of Artificial Intelligence ideas in a system which would rely heavily on good software and hardware engineering for achieving acceptable performance. The system needs to perform functions of signal processing, searching, pattern matching and syntactic and semantic analysis in close-to-real time. This can be done only with special purpose hardware. A set of parallel processors would operate on the digitised speech input and the constraining knowledge bases. Some of the processes, for example those dedicated to signal processing and pattern matching, would be implemented entirely in hardware.

The system described here is, despite all the detail given, a rather basic system. Ideally it should be expanded to include a text-to-speech system for flexible interaction with a user and a full word-processing system with business letter templates and so on.

REFERENCES

AHO, A.V. & CORASICK, M.J. (1975) "Fast pattern matching: An aid to bibliographic search", Comm. ACM, 18, 333-340.

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., GILLIS, J., ROSE, P. & WAGNER, M. (1986) "Deciphering Speech Waveforms". Proceedings of IEEE-IECEJ-ASJ International Conference on Acoustics, Speech, and Signal Processing, Tokyo, 2227-2230.

MEAD D. & O'KANE M. (1986) "A declarative approach to continuous speech recognition". Proceedings of the First Australian Conference on Speech Science and Technology, Canberra.

QUILLIAN, J.R. (1968) in Minsky, M., (ed.), "Semantic Information Processing", MIT Press, Cambridge, MA.