AUTOMATED SPEECH RECOGNITION: A FRAMEWORK FOR RESEARCH

Anne Johnstone
Department of Artificial Intelligence
Edinburgh University
Hope Park Square, Meadow Lane
Edinburgh EHB 9LL. (GB)

Gerry Altmann
Department of Linguistics
Edinburgh University
George Square
Edinburgh EHB 9LL. (GB)

ABSTRACT

This paper reflects the view that the decoding of speech, either by computer systems or people, must to a large extent be determined by the ways in which the speaker has encoded the information necessary for its comprehension. We therefore place great emphasis on the use of psycholinguistics as a tool for the construction of models essential to the characterisation of the speech understanding task.

We are primarily concerned with the interactions between the various levels at which a fragment of speech can be described (e.g. acoustic-phonetic, lexical, syntactic, etc), and the ways in which the knowledge bases associated with each of these "levels" contribute towards a final interpretation of an utterance. We propose to use the Chart Parser as a general computational framework for simulating such interactions, since its flexibility allows various models to be implemented and evaluated.

Within this general framework we discuss problems of information flow and search strategy in combining evidence across levels of description and across time, during the extension of an hypothesis. We stress the importance of both psychological and computational theory in developing a particular control strategy which could be implemented within the framework.

Introduction

The decoding of speech, either by computer systems or people, must to a large extent be determined by the ways in which the speaker has encoded the information necessary for its comprehension. Such a view is supported by a large body of experimental evidence concerning the ways in which various factors (eg. predictability from context) affect both the acoustic clarity with which a speaker pronounces an utterance, and the strategy the hearer appears to use in identifying it. The task of the computer system is to mimic, though preferably model, this strategy. In order to do so, one should presumably draw on both computational and psychological theories of process. Such a dual approach has been shown to be feasible, and indeed desirable, by research into early visual processing (eg. Marr 1976) which has shown that there can come a point when psychological and computational descriptions become barely distinguishable. This analogy with early visual processing is significant because central to the development of the vision research was the notion of 'modelling': one can argue that a significant difference between the so-called '4th Generation' and '5th Generation' technologies is that with the former ad-hoc algorithms are applied to often incomplete and unreliable data, while with 5th Generation systems, algorithms are devised by first constructing qualitative models suited to the task domain.

We propose to use psycholinguistics as a tool for the construction of models essential to the characterisation of the speech understanding task. We believe that this approach is essential to the development of automated speech recognition systems, and will also prove beneficial to psychological models of human speech processing, the majority of which are underdetermined from a computational point of view. Rumelhart and McClelland have recently adopted a similar approach to account for the major findings in the psychological literature on letter perception. By constructing a detailed computational model of the processes involved they were able to give an alternative description of the recognition of certain letter strings, which was supported by subsequent psycholinguistic experiments. Rumelhart and McClelland emphasise the point that their results were not predictable 'on paper', but were the outcome of considerable experimentation with the computational model.

Requirements of the Computational Framework

The experience of the ARPA speech project, which resulted in the design of a number of speech recognition systems, has demonstrated that the task of controlling the interactions between the knowledge bases which make up the system is at least as problematic as that of defining the knowledge bases. Major inadequacies in the systems developed during the ARPA project can be attributed to an early commitment in one or more areas of design which were not apparent until final testing and evaluation of the complete system began. An architecture is required, therefore, which will permit the development in parallel and relatively independently of component knowledge bases and methods of deploying them computationally. It should also permit the evaluation and testing of solutions with partially specified or simulated components. This will ensure that the design of one component will not influence unduly the design of any other component, possibly to the detriment of both. In addition, we should have the ability to determine the consequences of component design decisions by testing their contributions to the overall goals of speech recognition.

In order to fulfill these requirements we propose to use the active chart parser (e.g. Thompson & Ritchie, 1984). This was specifically designed as a flexible framework for the implementation (both serial and parallel) of different rule systems, and the evaluation of strategies for using these rule systems. It is described below in more detail.

The Computational Model

The problem in designing optimal control or search strategies lies in combining evidence across different levels of description (e.g. acoustic-phonetic, morpho-phonemic, syntactic, etc.), and across time during the extension of a hypothesis, such that promising interpretations are given priority and the right one wins. In this section we shall consider just a few of the issues concerning this flow of information.

Automated speech systems, in particular those implemented during the ARPA-SUR project, have been forced to confront the errorfull and ambiguous nature of speech, and to devise methods of controlling the very large search space of partial interpretations generated during processing. Although the problem was exacerbated by the poor performance of the acoustic-phonetic processing used in these systems, the experimental evidence suggests that the solution will not be found simply by improving techniques for low-level feature detection. The situation appears to be analogous to that of visual processing, where "significant" features may be absent. If present, their significance may also be open to a number of interpretations.

Combining evidence across different levels of description requires the specification of information flow between these levels. Within the psychological literature, there is a growing tendency away from "strong" (or "instructive") interactions towards "weak" (or "selective") interactions. With the latter, the only permissible flow of information involves the filtering out, by one component, of alternatives produced by other components (cf. Marslen-Wilson & Tyler, 1980; Crain & Steedman, 1982; Altmann & Steedman, forthcoming), so in hierarchical terms no component determines what is produced by any other component beneath it. A strong interaction, on the other hand, allows one component to direct, or guide, actively a second component in the pursuit of a particular hypothesis. Within the computational literature, weak interactions are also argued for on "aesthetic" grounds such as Marr's principles of modularity and least commitment (Marr, 1982).

The strongly interactive heterarchical and blackboard models implemented in HWIM and Hearsay II respectively have been criticised for the extremely complex control strategies which they required. Problems arise with the heterarchical model "because of the difficulties of generating each of the separate interfaces and, more importantly, because of the necessity of specifying the explicit control scheme." (Reddy & Erman, 1975). Similar problems arise with existing blackboard models. Their information flow allows strong top-down direction of components, resulting once again in highly complex control strategies. Hierarchical models have other problems, in that they allow too little interaction between the knowledge sources: within a strictly hierarchical system, one cannot "interleave" the processes associated with each different level of knowledge, and hence one cannot allow the very early filtering out by higher-level components of what might only be partial analyses at lower levels. This situation (considered disadvantageous for reasons of speed and efficiency) arises because of the lack of any common workspace over which the separate components can operate. There is, however, much to be said for hierarchical systems in terms of the relative

simplicity of the control strategies needed to manage them, a consideration which is fundamental to the design of any speech recognition system.

The model currently being developed embodies a weak hierarchical interaction, since this seems most promising on both psychological and computational grounds. Unlike existing hierarchical or associative models, it uses a uniform global data structure, a "chart". Associated with this structure is the active chart parser.

The active chart parser consists of the following:-

1) A uniform global data structure (the Chart), represents competing pathways through a search space, at different levels of description, and at different stages of analysis. Complete descriptions are marked by "inactive" paths, called edges, spanning temporally defined portions of the utterance. These inactive edges have pointers to the lower level descriptions which support them. Partial descriptions are marked by "active" edges which carry representations of the data needed to complete them. For example, a syntactic edge, such as a noun phrase, may span any complete descriptions that partially support it, such as a determiner or adjective. In addition, it will carry a description of the syntactic properties (e.g. noun) any inactive lexical edge must have to count both as additional evidence for this syntactic description and as justification for its extension or completion. The type and complexity of the descriptions are determined by the rule based knowledge systems used by the parser, and are not determined by the parser itself.

2) A multi-level task queueing structure (the Agenda), which is used to order the ways in which the descriptions will be extended, through time and level of abstraction, and thus to control the size and direction of the search space. This ordering on the agenda is controlled by specifically designed search strategies which determine the minimum amount of search compatible with a low rate of error in description. The power and flexibility of this approach in tackling complex system building tasks is well set out in Bobrow et al. 1976).

3) An algorithm which automatically schedules additions to the Chart onto the Agenda for subsequent processing wherever such extensions are possible. That is to say, whenever a description which is complete at some level (an inactive edge) can be used to extend a partial description at some higher level (an active edge). The knowledge bases define what extensions are possible, not the parser.

To summarize, the chart is used to represent and extend pathways, through time and level of abstraction, through a search space. Within the chart, there are different types of path corresponding to different levels of description, each of which is associated with a particular knowledge source. To the extent that knowledge specific rules specify what counts as constituent pathways at the different levels of abstraction, a hierarchical flow in information is maintained. The weak interaction arises because alternative pathways at one level of description can be filtered through attempts to build pathways at the next "higher" level. This model differs from straightforward hierarchical models, but resembles associative models, in that knowledge sources contribute to processing without each source necessarily corresponding to a distinct stage of analysis in the processing sequence.

Having sketched the construction of the search space we must now decide upon a strategy for exploring that space. Most current psychological theories appear to assume strict "left-to-right" processing, although this requires tackling stretches of sound immediatedly which are of poor acoustic quality, and which are relatively unconstrained by higher level knowledge. The majority of systems developed during the ARPA project found it necessary to use later occurring information to disambiguate earlier parts of an utterance. Moreover, there is psycholinguistic evidence that the "intelligibility" of a particular stretch of sound increases with additional evidence from later "rightward" stretches of sound (Pollack & Pickett, 1963; Warren & Warren, 1970). We propose to adopt a system using a form of left-to-right analysis which could approximate to the power of middle-out analysis (used in HWIM and Hearsay II) but without requiring the construction of distinct "islands" and with less computational expense. This more precise method of using "right-context effects" depends on the priority scores assigned to paths. Such scores can be thought of, for present purposes, as some measure of "goodness of fit". The score on a spanning pathway (that is, a pathway which spans other pathways "beneath" it) is determined by the scores on its constituents, and so is partly determined by scores towards its right-hand end. By virtue of affecting the "spanning score", a score on one sub-path can affect the probability that another sub-path to its left (as well as to its

right) will finally be chosen as the best description for the acoustic segment it represents. We will use psycholinguistic techniques to interrogate the "expert" (i.e. statistically reliable experiments with human listeners), in order to determine both when such leftwards flowing information is most often used for the disambiguation of poor quality areas, and what sets of paths it will affect. It will be extremely useful to know whether people regularly rely on information from the right to disambiguate preceding stretches of sound, or whether this happens only at the beginning of utterances as the HWIM strategy suggests. Pollack and Pickett claim that there is no effect on intelligibility of a word's position within a stimulus, but unfortunately they offer no inferential statistics to back this claim.

This is only one of the many issues in speech recognition which are experimentally addressable. The results of such experiments are obviously of relevance to computational systems since they can indicate where and when sources of information are most likely to contribute towards identification of an utterance. Conversely, the attempt to build a working model of at least some parts of the process, will highlight many areas where further experimental data is needed.

Concluding Remark

We hope that this sketch of part of the proposed system has given a feel for the combined approach taken here. It developed through a re-examination of a number of issues which arose during the ARPA speech project, and a reconsideration of these issues in the light of recent computational and psycholinguistic advances. Given the success of these recent advancements in the contributing fields of research, we feel that the time is right for the evaluation of a speech recognition system along the lines laid down here.

ACKNOWLEDGEMENTS

This is a summary of a paper (Johnstone & Altmann, 1984) written as a result of discussions held in the University of Edinburgh School of Epistemics research workshop on Lexical Access and Speech Perception. We would like to thank the members of that workshop, in particular Dr. Ellen Bard and Dr. Henry Thompson.

The proposals contained therein have been adopted by the Edinburgh contribution to the Plessey Consortium's Alvey Large Scale Demonstrator Project on Machine Assisted Speech Transcription.

REFERENCES

Altmann, G.T. & Steedman, M.J. Forthcoming. The Garden Path in Context: Reference and the Resolution of Local Syntactic Ambiguity.

Bard, E.G. & Anderson, A.H. 1983. The unintelligibility of speech to children. Journal of Child Language 10, 265-292

Crain, S. & Steedman, M.J. 1982. On not being led up the garden path: the use of context by the psychological parser. In (eds.) Dowty, D., Kartunnen, L., & Zwicky, A. Natural Language Parsing: psychological, computational, and theoretical perspectives. In press.

Johnstone, A.M. & Altmann, G.T. 1984. Automated Speech Recognition: A Framework for Research. Department of Artificial Intelligence, University of Edinburgh, Research Paper No. 233. Also appears as Speech Input Project, University of Edinburgh, Research Report No. 2.

Marr, D. 1976. Early Processing of Visual Information. Proc. Roy. Soc., 275.b

Marslen-Wilson, W.D. & Tyler, L.K. 1980. The Temporal Structure of Spoken Language Understanding. Cognition, 8, 1-71.

McClelland, J.L & Rumelhart, D.E. 1981. An Interactive Activation Model of Context Effects in Letter Perception: Part I. An Account of Basic Findings. In Psychological Review, 88, 375-407.

Pollack, I. & Pickett, J.M. 1963. The intelligibility of excerpts from Conversation. Language and Speech, 6, 165-171.

Reddy, D.R. & Ermann, L.D. 1975. Tutorial on System Organisation for Speech Understanding. In D.R. Reddy (ed) Speech Recognition, Academic Press.

Rumelhart, D.E. & McClelland, J.L. 1982. An Interactive Activation Model of Context Effects in Letter Perception: Part II. The Contextual Enhancement Effect. Some Tests and Extensions of the Model. In Psychological Review, 89, 60-94.

Thompson, H.S. & Ritchie, G.D. 1984. Techniques for Parsing Natural Language: Two Examples. In M. Eisenstadt & T. O'Shea (eds.) Artificial Intelligence Skills. Harper & Row.

Warren, R.M. & Warren, R.P. 1970. Auditory Confusions and Illusions. Scientific American, 223, 30-36.