SOME ISSUES IN KNOWLEDGE-BASED SPEECH RECOGNITION

John H. Connolly*

ABSTRACT

The knowledge-based approach to ASR gives rise to various AI-related questions, pertaining to the kinds of knowledge which ASR systems require, the capture and representation of that knowledge, and its use within such systems. These issues are discussed in the present paper, with a view to indicating the directions in which their resolution may lie.

INTRODUCTION

The need to incorporate linguistic knowledge into automatic speech recognition or understanding systems in order to make it possible for these systems to operate effectively has been pointed out by a number of investigators, for example Goodman & Reddy (ref 1), Hayes-Roth (ref 2). The notion of 'knowledge-based' speech recognition has arisen out of such conclusions, and represents an approach whereby the requisite knowledge is represented explicitly within the machine, and utilised directly by the system when evaluating possible hypotheses as to the linguistic composition of utterances spoken into it. This approach to automatic speech recognition (ASR) lies within the field of artificial intelligence (AI), and raises questions characteristic of knowledge engineering research, such as the following:

(1) (a) What kinds of knowledge should be incorporated?
(b) How should the knowledge be captured?
(c) How should it be represented and structured?
(d) How should it be used in the recognition process?

Each of these issues is addressed in the present paper, with particular reference to the design of research systems aimed at the speaker-independent recognition of continuous speech over large vocabularies.

KINDS OF KNOWLEDGE

It is generally accepted that both linguistic and extra-linguistic knowledge is pertinent to ASR, and that within the linguistic domain, knowledge relating to a variety of levels of structure is relevant. Indeed, information relating to each of the different structural levels can actually be crucial to successful speech recognition, as will now be demonstrated.

Beginning with segmental phonology, it is clear that unless the system can somehow detect the differences between one word and another, including minimal pairs such as tar and car, then it cannot be regarded as successful; and such inter-word distinctions are signalled via segmental phonology (in this case /t/ versus /kl/). Nonsegmental phonology is also significant to certain inter-word distinctions, such as aug'ust and 'August, which are differentiated through their rhythmic patterns; the latter would be crucial in the disambiguation of a (spoken) phrase like an aug'ust/August anniversary. Of course, the system will be able to distinguish between alternative words only to the extent that it is provided with an inventory of vocabulary items, so that lexical information, too, is indispensable to its operation. Moreover, in order to be speaker-independent, it needs to be provided with the knowledge required to cope with linguistic variation, such as differences in accent.

Grammatical knowledge is crucial in cases such as the following. Suppose that a person utters the sentence Choose one or the other and pronounces the conjunction as /a/, which is also a possible realisation of a, are, her or of. Here the system needs to be guided by grammatical considerations to select or rather than any of the alternatives, on the grounds that or is the only choice that results in a well-formed construction. On the other hand, it is semantics rather than grammar that represents the

*LUTCHI Research Centre, Department of Computer Studies, Loughborough University of Technology.
only means of identifying the appropriate lexical verb in sentences like *The shed was razed to the ground* compared with *The miners' cage was raised to the ground*. Pragmatic considerations, too, can be the vital factor in recognition, as for instance in the case of a sentence like *The mare /mayor was in excellent form*, where one would expect the noun functioning as subject to be identified differently according to whether the topic of conversation was horses or civic dignitaries. Consequently, the incorporation of linguistic knowledge into ASR systems is not a luxury but a necessity if those systems are not to be artificially constrained in terms of the input which they can be expected to process reliably.

With regard to the extra-linguistic knowledge that may be required, this is much more difficult to classify. The categorisation and handling of such knowledge with reference to ASR systems will, therefore, be left as a matter for future research.

**KNOWLEDGE CAPTURE**

If it is necessary to represent linguistic knowledge computationally within ASR systems, the issue arises of how to capture that knowledge in an appropriate form. Essentially, the activity of knowledge capture may be viewed as a three-stage process, though this should not be taken to imply that the whole of one stage has to be completed before the next begins, or that it might not be possible to collapse adjacent stages into one. The three stages are as follows:

(2) (a) Externalisation of the knowledge.
(b) Formalisation of the externalised knowledge.
(c) Encoding of the formalised knowledge into computer-usable form.

As Chomsky (ref 3) has pointed out, native speakers of a language possess a knowledge of that language, but this knowledge is not directly accessible to observation. In order to make the knowledge explicit it is necessary to employ the empirical techniques of descriptive linguistics, through which it is possible to arrive at a partial description of the language concerned. Such descriptions are conventionally couched in the technical language of linguistics, and for this reason (among others) they cannot be considered isomorphic with the native speaker's internalised knowledge. Rather, consisting as they do of statements about the language in question, they are more accurately regarded as encapsulations of the metalinguistic knowledge acquired by linguists. This knowledge is, therefore, a kind of expert knowledge, and in this sense an ASR system which embodies it can be seen as an expert system.

The next stage is to formalise the linguistic description into a coherent, explicit system. The type of system that results must be tailored to the level(s) to which it applies, but at the syntactic level, at least, it normally takes the form of a generative grammar of some kind. An advantage of casting grammatical knowledge in terms of a generative system is that it may be possible to convert the generative grammar into a computer program and test it in order to check that it really does generate what one expects it to (cf. Connolly, ref 4), as it is certainly possible to write grammars which do not! If the grammar can be tested in this way (and iteratively modified as necessary), and if, in addition, the computational representation of the grammar can be kept separate from the control structure of the testing program, then the encoding stage (2c) of the knowledge capture process for the ASR system will have been accomplished, and a higher level of confidence will be justified as to the reliability of the computational representation of the knowledge than if that representation had not been subjected to testing.

Special problems arise in connection with knowledge capture relating to the phonological level. Consider, for instance, the ASR system which is being developed at the LUTC.HI Research Centre, and which is based on automatic spectrogram analysis (see Johnson, Connolly & Edmonds, ref 5, Connolly et al., ref 6, Guzy & Edmonds, ref 7). Perhaps the chief difficulty here is that the shapes, such as formants, which humans are able to perceive in spectrograms are very hard to define with the rigour needed in order for them to be tractable as programming entities. It is, therefore, necessary to select, and iteratively refine, definitions relating to configurations of spectral energy which we can at least relate to our own percepts, and to adapt our spectral analysis techniques so as to make them operate in terms of these well-defined configurations.
However, despite the problems, there are advantages to be gained through working at the phonological level in terms of explicit knowledge encodings in association with representations of speech input that are immediately relatable to the constructs of phonetic theory. For example, if it is known that only the first two or three formants are relevant to vowel identification, then the system can exploit this knowledge so as to exclude the higher formants from consideration when performing the task in question. Moreover, operating with standard and familiar theoretical constructs is helpful when it comes to trying to find ways of improving the ASR system.

**REPRESENTATION AND STRUCTURE**

Whenever a body of knowledge is to be expressed in machine-usable form, it is necessary to consider exactly how the knowledge is to be represented and structured within the computer, so that it may be encoded accordingly. In the case of the knowledge required for an ASR system, there are grounds for representing that knowledge in the following manner:

1. In declarative form.
2. In a format permitting logical manipulation.
3. Within a modular structure.

Expressing the knowledge in declarative form carries with it certain advantages. It helps make for readability in the computer-encoded representation of the knowledge, thus making maintenance easier. Moreover, because it leads to the knowledge being separate from the program control structure, it means that statements referring to particular aspects of the encoded knowledge can be amended as required without necessarily affecting any of the procedural algorithms that may make use of it. This again makes for ease of maintenance, and it also facilitates the strategy of incorporating tested representations into the system, as described earlier. Representing the knowledge within a format that facilitates its manipulation by means of logical processes means that the declarative knowledge can be coupled with an inference engine in a manner which is reasonably well understood as a result of existing work in expert systems; see, for instance, Aly & Coombs (ref 8). Structuring the knowledge representation in terms of separate modules, with distinct knowledge sources for different linguistic levels, is a strategy which has been adopted in some of the ARPA SUR systems such as Hearsay-II and HWIM; see Erman & Lesser (ref 9), Wolf & Woods (ref 10). Modularity is also advocated by various writers on computational linguistics, such as Grishman (ref 11). The advantage is once again ease of maintenance, in that changes to one knowledge source do not inevitably impinge on any of the other modules. With regard to the computer language in which to encode the knowledge, Prolog is particularly suitable, since it is both declarative and designed to operate in terms of logical inference.

**USE IN THE RECOGNITION PROCESS**

The provision of machine-internal representations of knowledge is a necessary but not a sufficient precondition to the operation of a knowledge-based ASR system. The knowledge has also to be accessed in an suitable manner. Because an ASR system of the type under consideration works by evaluating possible hypotheses as to the linguistic composition of the input, it has to have the capacity to make decisions. Therefore, as proposed in Connolly et al. (ref 6), it is appropriate to access the knowledge by means of an inference system based upon the application of logical tests. To take a simple example, it is possible to formulate a test which has the effect of deciding whether or not there is evidence that a strident fricative has been articulated at some point in the input. If the result of the test is positive, then the system can infer that it should exclude from consideration any hypothesis that a word not containing such a fricative was being uttered at the moment in question. Thus, only words that do contain a strident should be accessed in the lexicon at this point in the processing of the input. Moreover, if the test is formulated as a binary (positive/negative) decision, in the manner just suggested, this has the advantage of permitting whole sets of unsustainable hypotheses (i.e. those which fail the test) to be completely eliminated, thereby reducing the range of hypotheses requiring further processing. If, on the other hand, the test were designed just to assign different statistical confidence ratings to the presence versus the absence of the strident, then unless the confidence rating for one or other outcome of the test were zero (an outcome which could not, of course be guaranteed), the non-preferred hypotheses would still have to be maintained in the system or else arbitrarily rejected if they fell below some threshold.
The use of binary logical tests thus facilitates a principled reduction of the search space through the systematic elimination of untenable hypotheses, making the recognition task correspondingly more manageable.

CONCLUSION

From the preceding discussion it may be concluded that (i) a knowledge-based ASR system requires components covering the whole range of linguistic levels; (ii) the capture of the relevant knowledge may be seen as a three-stage process of externalisation, formalisation and encoding; (iii) there is reason to represent the knowledge in a logically manipulable, declarative form, organised into a series of modules; and (iv) it is advantageous to access the knowledge representations through binary logical tests.

ACKNOWLEDGEMENT

In preparing to write this paper, the author has benefitted from discussion with a number of colleagues, particularly Julius Guzy.

REFERENCES

2. F Hayes-Roth Syntax, semantics and pragmatics in speech understanding systems (in Lea, ref 12) p 206.