Natural Language Understanding and Speech Recognition: Exploring the Connections

C. Rowles, X. Huang, and G. Aumann
Telecom Research Laboratories

ABSTRACT - This paper describes research aimed at integrating natural language understanding (NLU), speech recognition (SR) and the intonational structure of spoken language. NLU is being used to provide a measure of robustness to SR by placing utterances into context based on pragmatics and correctly recognize the speaker's intention in a database application. The approach is to use context to correct speech recognition errors and to reduce the search space. In turn, basic intonational structure derived from the speech waveform will assist lexical disambiguation, phrase attachment, and anaphoric resolution not dealt with by discourse segmentation. The paper outlines the speech understanding system architecture and describes the understanding process, giving examples of the use of intonation.

INTRODUCTION

The last few years have seen significant advances in speech recognition, the process of recognizing words from speech. The current state of the art uses Hidden Markov Modelling techniques to achieve of the order of 95% word recognition accuracy for speaker-independent, continuous-speech with moderately-sized vocabularies under laboratory conditions. While further research will certainly increase vocabulary size, larger vocabularies allow less structured and wider ranging interactions bringing new problems in speech recognition and the understanding of what the recognized words were meant to represent.

Limited vocabulary speech recognition systems are quite suitable for many applications where tasks are simple, interaction is structured and initiative can be seized by the understanding system. Here, specific words can relate complete intentions. Broaden the scope of interaction and increase the size of the vocabulary to make interaction more natural, however, and understanding speech becomes not just a problem of recognizing words, but also a problem of recognizing what those words represent.

There is a broad range of linguistic information that must be incorporated into a speech understanding system to support natural communications. For example, it has been long recognized that pragmatics are crucial in the understanding of natural language. Pragmatics include models of discourse and dialogue, domain semantics and the way people plan real-world activities. These things are often taken for granted to the extent that natural language expressions can be meaningless if not placed in the context of the conversational topic or conversation structure. In particular, pragmatics can assist in the understanding of ambiguity, reference and ellipsis.

Similarly, in spoken language, people use intonational cues such as amplitude, pitch and duration to indicate which interpretation of their utterance is the intended one. At the word recognition level, stress can assist in identifying words from phoneme strings, but it is at the semantic interpretation and discourse understanding levels that they prove to be of particular use for dealing with lexical ambiguity, phrase attachment, anaphoric resolution and the segmentation of discourse according to context boundaries.

The aim of our research is to explore pragmatics and intonational features of spoken language that can improve the semantic accuracy (that is, the accuracy with which the intended meaning of spoken discourse is recognized as distinct from the accuracy with which a string of words is correctly recognized) of speech recognition and bridge the gap between speech recognition and speech *understanding*. This implies recognizing a speaker's intentions, not just their utterances. The domain we have chosen, for demonstration purposes, is that of spoken access to an electronic directory assistance service. This narrows our domain primarily to that of information seeking requests, such as for names, phone numbers and so on.

This paper gives an overview of this project in progress. We do not attempt to give a complete description of how the intonational structure of spoken language or pragmatic information can be used to assist in the understanding of speech: we do not have all of the answers. Instead we will describe an architecture for an

experimental spoken language understanding system and show how aspects of intonation and pragmatics can be used to solve some of the difficult problems in the understanding process.

A SPEECH UNDERSTANDING SYSTEM: AN OVERVIEW

Our work is aimed at the construction of a prototype system for the understanding of spoken requests to an electronic directory assistance service. The proposed architecture of this system is shown in Figure . Our immediate work does not concentrate on the recognition of words from speech. Instead, we assume that we have available a string of word candidates that represents, possibly incompletely, the spoken input. Alternative word candidates are assumed to be available also. This Word Recognizer is supplemented by a Prosodic Feature Extractor, which produces features synchronized to the word string.

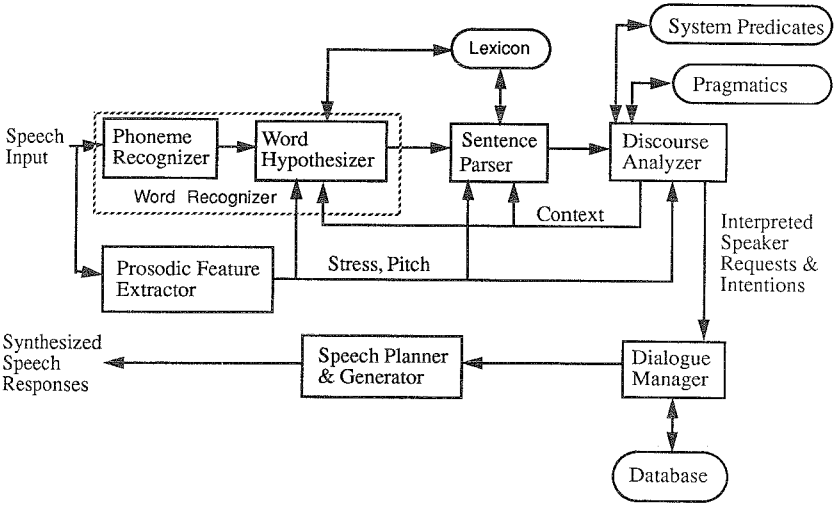

Figure 1. Architecture of a Speech Understanding System.

The output of the Word Recognizer is passed to a sentential-level parser (note: the term "sentence" is used loosely to mean a contiguous utterance of words). Prosodic features may be used by the parser for disambiguation, and context for selection of appropriate word senses. Sentence parse trees are then passed on to the Discourse Analyzer whose role is to segment the speaker's discourse into contextually consistent sub-discourses and interpret speaker requests in terms of available system functions. Prosodic cues assist in segmentation and classification of sub-discourses. A Dialogue Manager then integrates the speaker's discourse into a dialogue structure, manages interaction with the speaker and manges the retrieval of information from the database.

Work to date has centred on the Parser, Discourse Analyzer and Prosodic Feature Extractor. The paper will concentrate on these areas and some aspects of dialogue management. Speech planning and generation will also not be discussed.

THE PARSER

The Parser takes input from the Word Recognizer, in the form of a string of word candidates, plus prosodic information such as word stresses (phoneme stresses are not considered to be of great significance at this stage), pitch contours (or sentence tunes), and pauses.

The Parser works top-down and allows backtracking. It enforces sets of syntactic and semantic constraints on the input string in an attempt to reach an unambiguous interpretation of the input string. A presumption is made that the input string is well-formed. The presumption, however, is not a strong one, especially in the context of spoken dialogue processing. In reality the contrary is perhaps much more probable, either due to the speaker's "incorrect" utterance (people do not always speak grammatically), or due to the word recognizer's (WR) faulty processing result. The parser, therefore, can choose one of the two following actions upon its failed first attempt to parse.

1. Assuming the WR correctly recognizes the speaker's utterance, the parser loosens its own syntactic or/ and semantic constrains in a principled manner to analyze the speaker's ill-formed utterance.

2. Assuming the utterance is well-formed but the current candidate word string is a distorted one, the parser asks the WR for the next candidate.

A further complication exists where an ill-formed utterance gets further distorted by the WR. More exchanges between the parser and the WR will be needed for handling such cases; not surprisingly, it may very well be the case that other knowledge sources such as pragmatics, discourse analysis, and dialogue management will be called upon to help find the most likely interpretation for the input string.

Our strategy is a combination of actions 1 and 2: first, the system tries to obtain a parse for the current candidate word string, employing the parser's multi-level relaxation technique for handling ill-formed sentences (Huang &Chen, 1989); then, if no acceptable analysis is produced by action 1, the parser asks the WR to provide the next alternative word string.

During the parsing process prosodic information is used to help disambiguate certain structures which might be considered 'genuinely ambiguous' if no such prosodic information were available. For instance, the text form phrase:

old men and women in glasses

may produce four interpretations:

[old men] and [women in glasses]

[old men (in glasses)] and [(old) women in glasses]

[old men (in glasses)] and [women in glasses]

[old men] and [(old) women in glasses]

whereas in spoken form, with the help of prosodic information, in particular word stress, we might pinpoint a unique interpretation for the phrase.

For example, in processing the sentence

I ran into some old men and women in glasses walking down the street last night.

the prosodic information can help the system to decide on a unique interpretation, whereas without it the number of possible interpretations may reach more than eight (four interpretations for *some old men and women in glasses*,with at least two ways to attach *walking down the street*, either to *I* or to the conjoined noun phrase).

The output of the parser is composed of two parts, one is a parse tree and the other is discourse information obtained for the input string. The former contains syntactic, semantic and prosodic information. Most

ambiguity is removed in the parse tree, though some is left for later resolution, such as definite and anaphoric references, whose resolution normally requires inter-sentential inferences.

Discourse information includes both the discourse marker (which might be empty) accompanying the input string and the discourse type assigned to the string. Informally, discourse markers are cue words or word groups that link up propositions logically, temporally, or otherwise, such as *because, so, by the way, yes but,* etc. (cf. Reichman, 1985). Each input string is assigned a discourse type, based on the discourse marker found for the string; ambiguity may exist when no discourse marker is available, or when one discourse marker may indicate more than one discourse type. In such cases the ambiguity is passed on for later components to revolve. Currently we have defined, as a working set, twelve discourse types, namely opening, declarative, elaboration/clarification, correction, conversational move, semantic return, interruption, digression, confirm, reject, information request, and closing.

DISCOURSE ANALYSIS

The first stage of discourse analysis is the mapping of the parse tree into a *discourse schema.* Discourse schemas are structures consisting of a system-recognizable function, or (discourse) predicate, and a list of attributes that the function requires. The mapping selects a predicate based on such information as the primary verb, the mood and the modality of the parsed sentence, and then attempts to find values from the tree for each of the attributes in the schema. Not all attributes must have values instantiated as some will have defaults. The discourse schemas represent the system's view of the discourse purpose of the speaker's utterance. Discourse predicates include **inf_req** (information request such as a *wh* question), **decl_param** (declare a parameter value), **confirm** (a confirmation), **reject** (a rejection) and others appropriate to an electronic directory application.

Our approach to discourse analysis applies the focus tracking of Grosz and Sidner (Grosz &Sidner, 1987) and the discourse segmentation of Reichman (1978) to the incremental recognition of the speaker's plan. A full description of the discourse analysis is given in (Rowles, 1989), but it is appropriate to give a brief description here.

The role of discourse analysis is to take a collection of utterances from a speaker that may or may not be grammatically well-formed or complete, and that may have been incorrectly recognized by the speech recognition process, and produce a well-formed interpretation of the speaker's domain intention. That is, what it is that the speaker wishes to communicate to the system. Obviously, this interpretation is constrained by the capabilities of the system. By using pragmatic information, in particular domain semantics, possible user domain plans, and the structure of dialogues, we can place utterances into their domain or dialogue context and reach a more accurate interpretation of the speaker's intentions than by semantic interpretation alone. For example, discourse analysis can determine word senses, resolve pronominal and deictic reference and deal with complex dialogues that may include elaborations, interruptions and corrections. Eventually, the discourse analyzer will feed the context of a segment (the partially instantiated action schema) back to the parser and word hypothesizer to restrict the choice of words and word senses to those consistent with a segment context. Relaxation will be required to deal with significant conversational moves.

The Discourse Analyzer uses pragmatic information to aggregate partially or completely instantiated discourse schemas into contextually consistent discourse segments. The analyzer attempts to complete an *action schema* for each segment. An action schema represents a system capability (such as **retr_inf** , retrieving a specific piece of information) and is generated by recognizing the required action predicate and attribute values from the discourse schemas in a discourse segment. Each segment and hence, action schema, thus represents a specific intention or goal of the speaker. Segments may be hierarchically nested to reflect sub-dialogues such as interruptions, digressions and semantic returns.

The segmentation process uses linguistic and pragmatic information. For example, discourse segment boundaries may be indicated by the existence of cue words or phrases (e.g. *Now, I want to ...*), tense changes, context changes or intonational cues. The absence of these cues together with the presence of *continuation* sub-dialogues (for example, a clarification discourse schema following a declarative schema

with contextually consistent attribute values) indicates that the discourse schemas belong to the same segment. Thus, the following two sentences would interpreted as two distinct discourse schemas, but during discourse segmentation they would form the one discourse segment with just the one action schema:

I want a 9am booking at... => (declare !action book !time 9am ...)

No, make that 10am. => (correct !value 10am ...)

] action schema: (book !time 10am...)

Intonation is a useful cue here, as context changes may be difficult to detect, cue words usually occur at the start of a sentence (and so we may have to wait for the next one to detect that the last one completes a segment), and tense is only of limited use in our application. At present it is our aim to use the slope of the pitch contour over a sentence as a simple indicator (combined with other cues) of segment boundaries. For example, the final lowering of pitch in a declarative can indicate the completion of a topic while pitch rises towards the end of an interrogative (Hirschberg &Pierrehumbert, 1986). In addition, change in discourse context is often signalled by a rise in pitch. These intonational features are not absolute segment boundary indicators, however, and must be considered along with other indicators.

Of course many words that can act as discourse markers may have other, semantic, roles in a sentence, making the recognition of their discourse marker role difficult. Here again, intonational cues assist with difficult language understanding problems. Hirschberg and Litman (1987) have shown that cue words either form their own intonational phrase with clear boundary tone indication or begin an intonational phrase. Note that an intonational phrase may not directly correspond with a syntactic phrase.

An important role of discourse segmentation is to assist in the resolution of pronominal or deictic reference illustrated in the above discourse fragment as *that* referring to the booking time). In a discourse or dialogue, such reference problems are most commonly resolved by assuming that the correct referent is the most recent, contextually-consistent referent, either within the current discourse segment or related segments. Obviously, determining the correct referent may be difficult. Here a proword is commonly de-accented if this assumption is correct or accented if a different interpretation is required (Hirschberg & Pierrehumbert, 1986).

The work on using intonational cues in parsing and discourse analysis is in its early stages at present. We have identified some key cues that will greatly assist in certain natural language understanding problems, while the natural language understanding techniques solve some of the speech recognition difficulties. The stream of recognized words from the speech recognizer potentially contains errors, insertions and omissions. Occasional wrongly recognized words or words formed by incorrect phoneme string segmentation can be corrected by the parser or discourse analyzer. For example, an incorrect word input to the parser will be rejected if its grammatical role is inconsistent with the rest of the sentence and the word hypothesizer would have to backtrack and suggest another. On the other hand, if a word had a valid grammatical role and was important to the semantics of the sentence, it would be passed on to the discourse analyzer which would then reject it because of pragmatic inconsistence. The use of context by the word hypothesizer should reduceincorrect word hypotheses.

The prosodic features, primarily the pitch contour at this stage, are extracted from the speech waveform by dedicated signal processing hardware. This hardware provides some degree of speaker relative pitch independence, and currently has filtering aimed at determining the changes in pitch from word to word. It can, however, be easily changed to extract phoneme pitch if required at a later stage.

DIALOGUE MANAGEMENT

The role of the dialogue manager is to manage the overall interaction with the speaker. Initially, it would introduce the system in an attempt to provide the speaker with an appropriate mental model of the electronic directory and its capabilities. From there, its primary role is to ensure that the speaker provides sufficient information to enable a useful search of the directory. After generating an interpretation of the speaker's intention this is repeated to allow the speaker to make changes. If the action schema was incomplete and the speaker had completed their request, then the dialogue manager instigates a question to complete the

details required. Initial experiments have suggested that a difficulty here is in determining when the speaker has completed their request. If the action schema is incomplete for example, should the system wait for more input or generate a query to complete it. At this stage it appears that the final lowering of the pitch of utterances maybe useful in determining whether the speaker has completed the request or not.

DISCUSSION

In this paper we have sketched an architecture for a speech understanding system to allow spoken information seeking requests of an electronic directory. By integrating speech recognition, prosody and natural language understanding, we hope to improve the robustness and semantic accuracy of speech recognition in a restricted domain. We have described how semantics and pragmatics can be used to improve these aspects and how certain intonational cues in the spoken utterance can overcome problems in semantic disambiguation, proword reference and discourse segmentation.

Currently, work on the prosodic feature (pitch) extraction, parser and discourse analyzer is well underway. Future research is aimed at generating models of spoken dialogues and the intonational structure of dialogues and integrating these with the overall architecture.

ACKNOWLEDGEMENTS

The permission of the Executive General Manager, Research, Telecom Australia to publish the above paper is hereby acknowledged. The authors have benefited from discussions with Robin King, Julie Von Willer and Christian Matthiessen who are involved in further work on this project.

REFERENCES

Grosz,B.J. & Sidner, C.L., (1989), *Attention, Intentions, and the Structure of Discourse*, In Language and Artificial Intelligence, M. Nagao (Ed.), (Elsevier Science Publishers B.V., North-Holland).

Hirschberg, J. & Pierrehumbert,J., (1986), *The Intonational Structure of Discourse*, 24th Annual Meeting of the Association for Computational Linguistics.

Hirschberg, J. & Litman, D., (1987), *Now Let's Talk About **Now**: Identifying Cue Phrases Intonationally*, 25th Annual Meeting of the Association for Computational Linguistics.

Huang, X-M. & Chen, L. (1989), *Robust Natural Language Processing And Intelligent Language Tutoring*, AI and Creativity: 1st Australian Knowledge Engineering Program, Melbourne.

Reichman, R., (1978), *Conversational Coherency*, Cognitive Science 2 (4).

Reichman, R. (1985), *Getting Computers to Talk Like You and Me*, (Cambridge: The MIT Press).

Rowles, C.D. (1989), *Recognizing User Intentions from Natural language Expressions*, First Australia-Japan Joint Symposium on Natural Language Processing, 157-166.