Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 9983–9988, October 1995
Colloquium Paper

This paper was presented at a colloquium entitled "Human–Machine Communication by Voice," organized by Lawrence R. Rabiner, held by the National Academy of Sciences at The Arnold and Mabel Beckman Center in Irvine, CA, February 8–9, 1993.

Integration of speech with natural language understanding

ROBERT C. MOORE

Artificial Intelligence Center, SRI International, Menlo Park, CA 94025

ABSTRACT **The integration of speech recognition with natural language understanding raises issues of how to adapt natural language processing to the characteristics of spoken language; how to cope with errorful recognition output, including the use of natural language information to reduce recognition errors; and how to use information from the speech signal, beyond just the sequence of words, as an aid to understanding. This paper reviews current research addressing these questions in the Spoken Language Program sponsored by the Advanced Research Projects Agency (ARPA). I begin by reviewing some of the ways that spontaneous spoken language differs from standard written language and discuss methods of coping with the difficulties of spontaneous speech. I then look at how systems cope with errors in speech recognition and at attempts to use natural language information to reduce recognition errors. Finally, I discuss how prosodic information in the speech signal might be used to improve understanding.**

The goal of integrating speech recognition with natural language understanding is to produce spoken-language-understanding systems—that is, systems that take spoken language as their input and respond in an appropriate way depending on the meaning of the input. Since speech recognition (1) aims to transform speech into text, and natural-language-understanding systems (2) aim to understand text, it might seem that spoken-language-understanding systems could be created by the simple serial connection of a speech recognizer and a natural-language-understanding system. This naive approach is less than ideal for a number of reasons, the most important being the following:

- Spontaneous spoken language differs in a number of ways from standard written language, so that even if a speech recognizer were able to deliver a perfect transcription to a natural-language-understanding system, performance would still suffer if the natural language system were not adapted to the characteristics of spoken language.
- Current speech recognition systems are far from perfect transcribers of spoken language, which raises questions about how to make natural-language-understanding systems robust to recognition errors and whether higher overall performance can be achieved by a tighter integration of speech recognition and natural language understanding.
- Spoken language contains information that is not necessarily represented in written language, such as the distinctions between words that are pronounced differently but spelled the same, or syntactic and semantic information that is encoded prosodically in speech. In principle it should be possible to extract this information to solve certain understanding problems more easily using spoken input than using a simple textual transcription of that input.

This paper looks at how these issues are being addressed in current research in the ARPA Spoken Language Program.

COPING WITH SPONTANEOUS SPOKEN LANGUAGE

Language Phenomena in Spontaneous Speech

The participants in the ARPA Spoken Language Program have adopted the interpretation of requests for air travel information as a common task to measure progress in research on spoken language understanding. In support of this effort, over 15,000 utterances have been collected from subjects by using either a simulated or actual spoken language Air Travel Information System (ATIS). Interpreting these utterances requires dealing with a number of phenomena that one would not encounter often in dealing with linguists' examples or even real written texts.

Among the most common types of nonstandard utterances in the data are sentence fragments, sequences of fragments, or fragments combined with complete sentences:

six thirty a m from atlanta to san francisco what type of aircraft
on the delta flight number ninety eight what type of aircraft
i would like information on ground transportation city of boston between airport and downtown

A particular subclass of these utterances might be dubbed "afterthoughts." These consist of an otherwise well-formed sentence followed by a fragment that further restricts the initial request:

i'd like a return flight from denver to atlanta evening flights
i need the cost of a ticket going from denver to baltimore a first class ticket on united airlines
what kind of airplane goes from philadelphia to san francisco monday stopping in dallas in the afternoon first class flight

Another important group of nonstandard utterances can be classified as verbal repairs or self-corrections, in which the speaker intends that one or more words be replaced by subsequently uttered words. In the following examples, groups of words that are apparently intended for deletion are enclosed in brackets:

i'd like [to] a flight from washington [to] that stops in denver and goes on to san francisco
[do any of these flights] [do] are there any flights that arrive after five p m
can you give me information on all the flights [from san francisco no] from pittsburgh to san francisco on monday

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Some utterances involve use of metonymy, in which a word or phrase literally denoting one type of entity is used to refer to a related type of entity:

i need flight information between atlanta and boston
what is the flight number that leaves at twelve twenty p m
what delta leaves boston for atlanta

In the first two utterances, properties of flights are attributed to flight information and flight numbers; in the third, the name *delta* is used to refer to flights on Delta Airlines.

Some utterances that perhaps could be viewed as cases of metonymy might better be interpreted simply as slips of the tongue:

does delta aircraft fly d c tens

In this utterance *aircraft* has simply been substituted for *airlines*, perhaps because of the phonetic similarity between the two words and semantic priming from the information being requested.

Finally, some utterances are simply ungrammatical:

what kinds of ground transportation is available in dallas fort worth
okay what type of aircraft is used on a flight between san francisco to atlanta
what types of aircraft can i get a first class ticket from philadelphia to dallas
from those show me that serve lunch

The first example in this list is a case of lack of number agreement between subject and verb; the subject is plural, the verb singular. The second example seems to confuse two different ways of expressing the same constraint; *between san francisco and atlanta* and *from san francisco to atlanta* have been combined to produce *between san francisco to atlanta.* The final pair of examples both seem to involve deletion of function words needed to make the utterances grammatical:

what types of aircraft can i get a first class ticket from philadelphia to dallas (on)
from those show me (the ones) that serve lunch

Of course, there are also utterances that combine several of these phenomena, for example:

[flight number] [okay flight] [let us] you have a flight number going [to] from san francisco to atlanta around eight a m

This utterance appears to begin with three repairs; it finally gets going with a somewhat odd way of asking for a flight number, *you have a flight number*. . . , it involves the metonymy of talking about a flight number going somewhere rather than a flight and includes yet another repair, replacing *to* with *from*.

Strategies for Handling Spontaneous Speech Phenomena

Spoken-language-understanding systems use various strategies to deal with the nonstandard language found in spontaneous speech. Many of these phenomena, although regarded as nonstandard, are just as regular in their patterns of use as standard language, so they can be incorporated into the linguistic rules of a natural language system. For example, one may add grammar rules to allow a sequence of syntactically unconnected noun phrase modifiers to be a complete utterance if all the modifiers can be interpreted as predications of the same class of objects, so that an utterance like *from boston to dallas on tuesday* can be interpreted as a request for flights or fares. Nonstandard but regular uses of particular lexical items can be accommodated simply by extending the lexicon. For example, common cases of metonymy can be handled this way; *flight information* and *flight number* can be lexically coded to allow the same modifiers as *flight*.

Extending the linguistic rules of a system to include nonstandard but regular patterns still leaves disfluencies and truly novel uses of language unaccounted for. To deal with these, virtually all systems developed for the ARPA ATIS task incorporate some sort of language-understanding method that does not depend on deriving a complete analysis of the input that accounts for every word. Such methods are usually described as "robust," although they are actually robust only along certain dimensions and not others. A common strategy is to have predefined patterns (case frames or templates) for the most common sorts of queries in the ATIS task and to scan the input string for words or phrases that match the elements of the pattern, allowing other words in the utterance to be skipped. The Carnegie–Mellon University (CMU) Phoenix system (3) and SRI International's Template Matcher (4) rely exclusively on this approach. In the SRI system, for example, a key word such as *flight*, *fly*, *go*, or *travel* is used to trigger the template for a request for flight information and phrases matching patterns such as *on* <date>, *from* <city>, and *to* <city> are used to fill in constraints on the flights. This allows the system to ignore disfluencies or novel language if they do not occur in parts of the utterances that are crucial for recognizing the type of request or important constraints on the request. For instance, this approach can easily process the example given above of a sentence with multiple problematic features,

[flight number] [okay flight] [let us] you have a flight number going [to] from san francisco to atlanta around eight a m

because it is fundamentally a very simple type of request, and none of the disfluencies affect the phrases that express the constraints on the answer.

This template-based approach works well for the vast majority of utterances that actually occur in the ATIS data. In principle, however, the approach would have difficulties with utterances that express more complex relationships among entities or that involve long-distance dependencies, such as in

what cities does united fly to from san francisco

Here the separation of *what cities* and *to* would make this utterance difficult to interpret by template-based techniques, unless a very specific pattern were included to link these together. But in fact this is a made-up example, and things like this are extremely rare, if they occur at all, in the ATIS task. Nevertheless, such possibilities have led several groups to design systems that first try to carry out a complete linguistic analysis of the input, falling back on robust processing techniques only if the complete analysis fails. The Delphi system of BBN Systems and Technologies (5), the TINA system of the Massachusetts Institute of Technology (MIT) (6), and SRI International's Gemini+TM system [a combination of the Template Matcher with SRI's Gemini system (7)] all work this way. In the case of SRI's systems, the combination of detailed linguistic analysis and robust processing seems to perform better than robust processing alone, with the combined Gemini+TM system having about four points better performance than the Template Matcher system alone for both speech and text input in the November 1992 ATIS evaluation, according to the weighted understanding error metric (8).* It should be noted, however, that the best-performing system in the November 1992 ATIS evaluation, the CMU Phoenix system, uses only robust interpretation methods with no attempt to account for every word of an utterance.

The robust processing strategies discussed above are fairly general and are not specifically targeted at any particular form

*This measure is equal to the percentage of utterances correctly answered minus the percentage incorrectly answered, with utterances not answered omitted, so as to punish a wrong answer more severely than not answering it all.

of disfluency. There has been recent work, however, aimed specifically at the detection and correction of verbal repairs. Utterances containing repairs constitute about 6% of the ATIS corpus, although repair rates as high as 34% have been reported for certain types of human–human dialogue (9). A module to detect and correct repairs has been incorporated into SRI's Gemini system (7, 10) that is keyed to particular word patterns that often signal a repair. For example, in the utterance

> *can you give me information on all the flights [from san francisco no] from pittsburgh to san francisco on monday*

the section *from san francisco no from pittsburgh* matches a pattern of a cue word, *no*, followed by a word (*from*) that is a repetition of an earlier word. Often, as in this case, this pattern indicates that text from the first occurrence of the repeated word through the cue word should be deleted. This kind of pattern matching alone generates many false positives, so in the Gemini system a repair edit based on pattern matching is accepted only if it converts an utterance that cannot be fully parsed and interpreted into one that can. Applying this method to a training corpus of 5873 transcribed utterances, Gemini correctly identified 89 of the 178 utterances that contained repairs consisting of more than just a word fragment. Of these, 81 were repaired correctly and 8 incorrectly. An additional 15 utterances were misidentified as containing repairs. Similar results were obtained in a fair test on transcriptions of the November 1992 ATIS test set. Gemini identified 11 of the 26 repair utterances out of 756 interpretable utterances, of which 8 were repaired correctly and 3 incorrectly; 3 other utterances were misidentified as containing repairs.

It should be noted that the general robust processing methods discussed above are insensitive to repairs if they occur in sections of the utterance that are not critical to filling slots in the pattern being matched. In addition, CMU's Phoenix system incorporates one other simple method for handling repairs. If the pattern matcher finds more than one filler for the same slot, it uses the last one, on the assumption that the fillers found earlier have been replaced by repairs. This method seems to work on many of the repairs actually found in the ATIS corpus. It is easy to make up cases where this would fail, but the more complex method used by Gemini would work:

> *show me flights to boston no from boston*

However, it is not clear that such utterances occur with any frequency in the ATIS task.

ROBUSTNESS TO RECOGNITION ERRORS

Since even the best speech recognition systems make at least some errors in a substantial proportion of utterances, coping with speech recognition errors is one of the major challenges for correctly interpreting spoken language. This problem can be approached in a number of ways. If there are particular types of errors that the recognizer is especially prone to make, the natural-language-understanding system can be modified to accommodate them. For instance, *four* can be allowed in place of *for*, or the deletion of short, frequently reduced words, such as *to*, can be permitted.

For the most part, however, current systems rely on their general robust understanding capability to cope with noncritical recognition errors. Although it has not been carefully measured, anecdotally it appears that a high proportion of recognition errors made by the better-performing recognizers for the ATIS task occur in portions of the utterance that are noncritical for robust interpretation methods. It may be conjectured that this is due to the fact that most of the critical key words and phrases are very common in the training data for the task and are therefore well modeled both acoustically and in the statistical language models used by the recognizers.

Table 1. Comparison of understanding error with recognition error in November 1992 ATS evaluation

System	Understanding error w/text, %	Understanding error w/speech, %	Recognition error, %
BBN	15.0	19.0	25.2
CMU	6.5	11.2	28.9
MIT	10.9	19.2	37.8
SRI	15.2	21.6	33.8

The degree of robustness of current ATIS systems to speech recognition errors can be seen by examining Table 1. This table compares three different error rates in the November 1992 evaluation (8) of the ATIS systems developed by the principal ARPA contractors working on the ATIS task: BBN Systems and Technologies (BBN), Carnegie–Mellon University (CMU), the Massachusetts Institute of Technology (MIT), and SRI International (SRI). The error rates compared are (*i*) the percentage of queries not correctly answered when the natural language component of the system was presented with verified transcriptions of the test utterances, (*ii*) the percentage of queries not correctly answered when the combined speech recognition and natural language understanding system was presented with the digitized speech signal for the same utterances, and (*iii*) the percentage of queries for which the speech recognition component of the system made at least one word recognition error in transcribing the utterance. All of these error rates are for the subset of utterances in the test set that were deemed to constitute answerable queries.

The striking thing about these results is that for all of these systems the increase in understanding error going from text input to speech input is surprisingly small in comparison with the rate of utterance recognition error. For instance, for the CMU system the rate of understanding error increased by only 4.7% of all utterances when a verified transcription of the test utterances was replaced by speech recognizer output, even though 28.9% of the recognizer outputs contained at least one word recognition error. Moreover, the rate of speech-understanding errors was much lower than the rate of speech recognition errors for all systems, even though all systems had many language-understanding errors even when pro-vided with verified transcriptions. This shows a remarkable degree of robustness to recognition errors using methods that were primarily designed to cope with difficult, but accurately transcribed, spoken language.

NATURAL LANGUAGE CONSTRAINTS IN RECOGNITION

Models for Integration

Despite the surprising degree of robustness of current ATIS systems in coping with speech recognition errors, Table 1 also reveals that rates for understanding errors are still substantially higher with speech input than with text input, ranging from 1.26 to 1.76 times higher, depending on the system. One possible way to try to close this gap is to use information from the natural language processor as an additional source of constraint for the speech recognizer. Until recently, most attempts to do this have followed what might be called "the standard model":

> Pick as the preferred hypothesis the string with the highest recognition score that can be completely parsed and interpreted by the natural language processor.

This model was embodied in several systems developed under the original ARPA Speech Understanding Program of the 1970s (11) and also in some of the initial research in the current

ARPA Spoken Language Program (12–15). However, the model depends on having a natural language grammar that accurately models the speech being recognized. For the kind of messy, ill-formed spoken language presented here in the section "Language Phenomena in Spontaneous Speech," this presents a serious problem. It is highly unlikely that any conventional grammar could be devised that would cover literally everything a person might actually utter.

The dilemma can be seen in terms of the kind of natural language system, discussed in the section "Strategies for Handling Spontaneous Speech Phenomena," that first attempts a complete linguistic analysis of the input and falls back on robust processing methods if that fails. If the grammar used in attempting the complete linguistic analysis is incorporated into the speech recognizer according to the standard model, the recognizer will be overconstrained and the robust processor will never be invoked because only recognition hypotheses that can be completely analyzed linguistically will ever be selected. This means that, for cases in which the robust processor should have been used, the correct recognition hypothesis will not even be considered. On the other hand, if the robust language processor were incorporated into the speech recognizer according to the standard model, it would provide very little information since it is designed to try to make sense out of almost any word string.

A number of modifications of the standard model have been proposed to deal with this problem. One method is to use a highly constraining grammar according to the standard model but to limit the number of recognition hypotheses the grammar is allowed to consider. The idea is that, if none of the top *N* hypotheses produced by the recognizer can successfully be analyzed by the grammar, it is likely that the correct recognition hypothesis is not allowed by the grammar, and a second attempt should be made to interpret the recog-nizer output using robust processing. The parameter *N* can be empirically estimated to give optimal results for a particular grammar on a particular task. Another possibility is to allow parsing a hypothesis as a sequence of grammatical fragments with a scoring metric that rewards hypotheses that can be analyzed using fewer fragments.

An additional problem with the standard model is that it does not take into account relative likelihoods of different hypotheses. All utterances that are allowed by the grammar are treated as equally probable. This differs from the *N*-gram statistical language models commonly used in speech recognition that estimate the probability of a given word at a particular point in the utterance based on the one or two immediately preceding words. Baker (16) developed an automatic training method, the inside-outside algorithm, that allows such techniques to be extended to probabilistic context-free grammars. Since most grammars used in natural-language-processing systems are based to some extent on context-free grammars, Baker's or related methods may turn out to be useful for developing probabilistic natural language grammars for use in speech recognition. This approach appears to leave at least two important problems to be addressed, however.

First, while the grammars used in natural-language-processing systems are usually based on the context-free grammars, they also usually have augmentations that go beyond simple context-free grammars. Recently, grammars based on the unification of grammatical categories incorporating features-value structures (17) have been widely used. If the value spaces are finite for all the features used in a particular grammar, the grammar is formally equivalent to a context-free grammar, but for any realistic unification grammar for natural language the corresponding context-free grammar would be so enormous (due to all the possible combinations of feature values that would have to be considered) that it is extremely doubtful that enough training data could either be obtained or processed to provide reliable models via the inside-outside algorithm. This suggests that, at best, a carefully selected context-free approximation to the full grammar would have to be constructed.

A second problem derives from the observation (18) that particular lexical associations, such as subject-verb, verb-object, or head-modifier, appear to be a much more powerful source of constraint in natural language than the more abstract syntactic patterns typically represented in natural language grammars. Thus, in predicting the likelihood of a combination such as *departure time*, one cannot expect to have much success by estimating the probability of such noun-noun combinations, independently of what the nouns are, and combining that with context-independent estimates of an arbitrary noun being *departure* or *time*. Yet in the model of probabilistic context-free grammar to which the inside-outside algorithm applies, this is precisely what will happen, unless the grammar is carefully designed to do otherwise. If probabilistic models of natural language are not constructed in such a way that lexical association probabilities are captured, those models will likely be of little benefit in improving recognition accuracy.

Architectures for Integration

Whatever model is used for integration of natural language constraints into speech recognition, a potentially serious search problem must be addressed. The speech recognizer can no longer simply find the best acoustic hypothesis; it must keep track of a set of acoustic hypotheses for the natural language processor to consider. The natural language processor similarly has to consider multiple recognition hypotheses, rather than a single determinate input string. Over the past 5 years, three principal integration architectures for coping with this search problem have been explored within the ARPA Spoken Language Program: word lattice parsing, dynamic grammar networks, and *N*-best filtering or rescoring.

Word Lattice Parsing

Word lattice parsing was explored by BBN (12, 13) in the early stages of the current ARPA effort. In this approach the recognizer produces a set of word hypotheses, with an acoustic score for each potential pair of start and end points for each possible word. A natural language parser is then used to find the grammatical utterance spanning the input signal that has the highest acoustic score. Word lattice parsing incorporates natural language constraints in recognition according to the standard model, but it results in extremely long processing times, at least in recent implementations. The problem is that the parser must deal with a word lattice containing thousands of word hypotheses rather than a string of just a few words. More particularly, the parser must deal with a large degree of word boundary uncertainty. Normally, a word lattice of adequate size for accurate recognition will contain dozens of instances of the same word with slightly different start and end points. A word lattice parser must treat these, at least to some extent, as distinct hypotheses. One approach to this problem (13) is to associate with each word or phrase a set of triples of start points, end points, and scores. Each possible parsing step is then performed only once, but a dynamic programming procedure must also be performed to compute the best score for the resulting phrase for each possible combination of start and end points for the phrase.

Dynamic Grammar Networks

Dynamic grammar networks (14, 15) were developed to address the computational burden in word lattice parsing posed by the need for the parser to deal with acoustic scores and multiple possible word start and end points. In this approach a natural language parser is used to incrementally generate the grammar-state-transition table used in the standard hidden Markov model (HMM) speech recognition architecture. In an HMM speech recognizer, a finite-state grammar is used to predict what words can start in a particular recognition state and to specify what recognition state the system should go into when a particular word is recognized in a given predecessor state. Dynamic programming is used to efficiently assign a score to each recognition state the system may be in at a particular point in the signal.

In HMM systems the finite-state grammar is represented as a set of state-word-state transitions. Any type of linguistic constraints can, in fact, be represented as such a set, but for a nontrivial natural language grammar the set will be infinite. The dynamic-grammar-network approach computes the state-transition table needed by the HMM system incrementally, generating just the portion necessary to guide the pruned search carried out by the recognizer for a particular utterance. When a word is successfully recognized beginning in a given grammar state, the recognizer sends the word and the state it started in to the natural language parser, which returns the successor state. To the parser, such a state encodes a parser configuration. When the parser receives a state-word pair from the recognizer, it looks up the configuration corresponding to the state, advances that configuration by the word, creates a name for the new configuration, and passes back that name to the recognizer as the name of the successor state. If it is impossible, according to the grammar, for the word to occur in the initial parser configuration, the parser sends back an error message to the recognizer, and the corresponding recognition hypothesis is pruned out. Word boundary uncertainty in the recognizer means that the same word starting in the same state can end at many different points in the signal, but the recognizer has to communicate with the parser only once for each state-word pair. Because of this, the parser does not have to consider either acoustic scores or particular start and end points for possible words, those factors being confined to the recognizer.

The dynamic-grammar-net approach succeeds in removing consideration of acoustic scores and word start and end points from the parser, but it too has limitations. The state space tends to branch from left to right, so if a group of utterance hypotheses differ in their initial words but have a later substring in common, that substring will be analyzed multiple times independently, since it arises in different parsing configurations. Also, in the form presented here, the dynamic-grammar-net approach is tied very much to the standard model of speech and natural language integration. It is not immediately clear how to generalize it so as not to overconstrain the recognizer in cases where the utterance falls outside the grammar.

N-Best Filtering or Rescoring

N-best filtering and rescoring were originally proposed by BBN (19, 20). This is a very simple integration architecture in which the recognizer enumerates the *N*-best full recognition hypotheses, which the natural language processor then selects from. The standard model of speech and natural language integration can be implemented by *N*-best filtering, in which the recognizer simply produces an ordered list of hypotheses, and the natural language processor chooses the first one on the list that can be completely parsed and interpreted. More sophisticated models can be implemented by *N*-best rescoring, in which the recognition score for each of the *N*-best recognition hypotheses is combined with a score from the natural language processor, and the hypothesis with the best overall score is selected.

The advantage of the *N*-best approach is its simplicity. The disadvantage is that it seems impractical for large values of *N*. The computational cost of the best-known method for exact enumeration of the *N*-best recognition hypotheses (19) increases linearly with *N*; but an approximate method exists (20) that increases the computational cost of recognition only by a small constant factor independent of *N*. There is no reported method, however, for carrying out the required natural language processing in time less than linear in the number of hypotheses. In practice, there seem to be no experiments that have reported using values of *N* greater than 100, and the only near-real-time demonstrations of systems based on the approach have limited *N* to 5. To put this in perspective, in information theoretic terms, an *N*-best system that selects a single hypotheses from a set of 64 hypotheses would be providing at most 6 bits of information per utterance. On the other hand, it has not proved difficult to develop purely statistical language models for particular tasks that provide 6 bits or more of information per *word*.[†]

However, if the basic recognizer is good enough that there is a very high probability of the correct hypothesis being in that set of 64, those 6 bits per utterance may be enough to make a practical difference.

Integration Results

In 1992, BBN (21) and MIT (22) reported results on the integration of natural language constraints into speech recognition for the ATIS task. BBN used an *N*-best integration scheme in which strict grammatical parsing was first attempted on the top five recognition hypotheses, choosing the hypothesis with the best recognition score that could be fully interpreted. If none of those hypotheses was fully interpretable, the process was repeated using a robust processing strategy. On a development test set, BBN found that this reduced the weighted understanding error from 64.6%, when only the single best recognizer output was considered, to 56.6%—a 12.4% reduction in the error rate. Taking the top 20 hypotheses instead of the top five also improved performance compared with taking only the single top hypothesis, but the improvement was less than when consideration was limited to the top five.

The MIT experiment was more complex, because it used natural language constraints in two different ways. First, a probabilistic LR parser was integrated into the recognition search (using an architecture that somewhat resembled dynamic grammar nets) to incorporate a language probability into the overall recognition score. The LR parser was modified to allow parsing an utterance as a sequence of fragments if no complete parse allowed by the grammar could be found. Then the top 40 recognition hypotheses were filtered using the complete natural language system. This reduced the word recognition error from 20.6 to 18.8% (an 8.7% reduction) and the utterance recognition error from 67.7 to 58.1% (a 13.4% reduction) on a development test set, compared with the best version of their recognizer incorporating no natural language constraints.

SPEECH CONSTRAINTS IN NATURAL LANGUAGE UNDERSTANDING

While most efforts toward integration of speech and natural language processing have focused on the use of natural language constraints to improve recognition, there is much information in spoken language beyond the simple word sequences produced by current recognizers that would be useful for interpreting utterances if it could be made available. Prosodic information in the speech signal can have important effects on utterance meaning. For example, since in the ATIS

[†] For a 1000-word recognition task, a perplexity-15 language model reduces the effective vocabulary size by a factor of about 67, which is about a 6-bit per word reduction in entropy.

task the letters B, H, and BH are all distinct fare codes, the two utterances

What do the fare codes BH and K mean?
What do the fare codes B, H, and K mean?

would differ only prosodically but have very different meanings.

In a preliminary study of the use of prosodic information in natural language processing, Bear and Price (23) reported on an experiment in which the incorporation of prosodic information into parsing reduced syntactic ambiguity by 23% on a set of prosodically annotated sentences. Another area where information from the speech signal would undoubtedly be useful in natural language processing is in the detection of verbal repairs. While no experiment has yet been reported that actually used speech information to help correct repairs, Bear *et al.* (10) and Nakatani and Hirschberg (24) have identified a number of acoustic cues that may be useful in locating repairs.

CONCLUSIONS

Work on the ATIS task within the ARPA Spoken Language Program is ongoing, but a number of conclusions can be drawn on the basis of what has been accomplished so far. Perhaps the most significant conclusion is that natural-language-understanding systems for the ATIS task have proved surprisingly robust to recognition errors. It might have been thought a priori that spoken language utterance understanding would be significantly worse than utterance recognition, since recognition errors would be compounded by understanding errors that would occur even when the recognition was perfect. The result has been quite different, however, with the robustness of the natural language systems to recognition errors more than offsetting language-understanding errors.

Nevertheless, understanding error remains 20 to 70% higher with speech input than with text input. Attempts to integrate natural language constraints into recognition have produced only modest results so far, improving performance by only about 9 to 13 percent, depending on how performance is measured.

Is it possible to do better? So far, relatively few ideas for the incorporation of natural language constraints into recognition of spontaneous speech have been tested, and there may be many ways in which the approaches might be improved. For example, published reports of experiments conducted so far do not make it clear whether strong semantic constraints were used or how well important word associations were modeled. Compared with where the field stood only 3 or 4 years ago, however, great progress has certainly been made, and there seems no reason to believe it will not continue.

Preparation of this paper was supported by the Advanced Research Projects Agency under Contract No. N00014-93-C-0142 with the Office of Naval Research.

1. Makhoul, J. & Schwartz, R. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 9956–9963.
2. Bates, M. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 9977–9982.
3. Ward, W., *et al.* (1992) in *Proceedings of the Speech and Natural Language Workshop, Harriman, NY* (Kaufmann, San Mateo, CA), pp. 78–83.
4. Jackson, E., *et al.* (1991) in *Proceedings of the Speech and Natural Language Workshop, Pacific Grove, CA* (Kaufmann, San Mateo, CA), pp. 190–194.
5. Stallard, D. & Bobrow, R. (1992) in *Proceedings of the Speech and Natural Language Workshop, Harriman, NY* (Kaufmann, San Mateo, CA), pp. 305–310.
6. Seneff, S. (1992) in *Proceedings of the Speech and Natural Language Workshop, Harriman, NY* (Kaufmann, San Mateo, CA), pp. 299–304.
7. Dowding, J., *et al.* (1993) in *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics* (Columbus, OH), pp. 54–61.
8. Pallet, D. S., *et al.* (1993) in *Proceedings of the ARPA Workshop on Human Language Technology, Plainsboro, NJ* (Kaufmann, San Mateo, CA), pp. 7–18.
9. Levelt, W. (1983) *Cognition* **14,** 41–104.
10. Bear, J., Dowding, J. & Shriberg, E. (1992) in *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics* (Newark, NJ), pp. 56–63.
11. Klatt, D. H. (1977) *Acoust. Soc. Am.* **62,** 1345–1366.
12. Boisen, S., *et al.* (1989) in *Proceedings of the Speech and Natural Language Workshop, Philadelphia* (Kaufmann, San Mateo, CA), pp. 106–111.
13. Chow, Y. L. & Roukos, S. (1989) in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing* (Glasgow, Scotland), pp. 727–730.
14. Moore, R., Pereira, F. & Murveit, H. (1989) in *Proceedings of the Speech and Natural Language Workshop, Philadelphia* (Kaufmann, San Mateo, CA), pp. 243–247.
15. Murveit, H. & Moore, R. (1990) in *Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing* (Albuquerque, NM), pp. 573–576.
16. Baker, J. (1979) in *Speech Communication Papers Presented at the 97th Meeting of the Acoustical Society of America*, eds. Wolf, J. J. & Klatt, D. H. (Massachusetts Institute of Technology, Cambridge, MA).
17. Shieber, S. M. (1986) *An Introduction to Unification-Based Approaches to Grammar* (Center for the Study of Language and Information, Stanford Univ., Stanford, CA).
18. Church, K., *et al.* (1989) in *Proceedings of the Speech and Natural Language Workshop, Cape Cod, MA* (Kaufmann, San Mateo, CA), pp. 75–81.
19. Chow, Y. L. & Schwartz, R. (1989) in *Proceedings of the Speech and Natural Language Workshop, Cape Cod, MA* (Kaufmann, San Mateo, CA), pp. 199–202.
20. Schwartz, R. & Austin, S. (1990) in *Proceedings of the Speech and Natural Language Workshop, Hidden Valley, PA* (Kaufmann, San Mateo, CA), pp. 6–11.
21. Kubala, F., *et al.* (1992) in *Proceedings of the Speech and Natural Language Workshop, Harriman, NY* (Kaufmann, San Mateo, CA), pp. 72–77.
22. Zue, V., *et al.* (1992) in *Proceedings of the Speech and Natural Language Workshop, Harriman, NY* (Kaufmann, San Mateo, CA), pp. 84–88.
23. Bear, J. & Price, P. (1990) in *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics* (Pittsburgh), pp. 17–22.
24. Nakatani, C. & Hirschberg, J. (1993) in *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics* (Columbus, OH), pp. 46–53.