Combining Linguistic with Statistical Methods in Automatic Speech Understanding

Patti Price
SRI International
333 Ravenswood Avenue
Menlo Park, California 94025
pprice@speech.sri.com

Introduction

This paper presents an overview of automatic speech understanding techniques that combine knowledge-based approaches with statistical pattern matching methods. Such an approach requires a multidisciplinary outlook, addressing both cultural and technical differences among the various component technologies.

As argued in Price and Ostendorf (1994), the representatives of knowledge-based approaches and of approaches based on statistical pattern matching may view each other with suspicion — if they are aware of each other's work. Psychologists and linguists, representing the knowledge-based approaches, may view automatic algorithms as "uninteresting collections of ad hoc ungeneralizable methods for limited domains." The automatic speech recognition community, on the other hand, may argue that automatic speech recognition should not be modeled after human speech recognition; since the tasks and goals of machines are very different from those of humans, the methods should also be different. Thus, in this view, knowledge-based approaches are "uninteresting collections of ad hoc ungeneralizable methods for limited domains." The two sides may use the same words, but mean different things, as indicated in the glossary in the table.

Spoken language is a *social* mechanism evolved for communication among entities whose biological properties constrain the possibilities. Therefore the mechanisms that are successful for machines are likely to share many properties with those successful for people. Further, in automatic spoken language applications, at least one human being is typically involved. Thus, the understanding of human communication may be essential for generalizable methods robust to the variability manifested by humans. Just as engineers could gain from a better understanding of human mechanisms, psychologists and linguists could gain from a better understanding of automatic techniques. For example, these techniques can be viewed as theories of human communication made explicit enough to test. Studying where the techniques work and where they fail could shed light on the human communication process. Although differences in training, techniques, approaches, goals, and culture have inhibited multidisciplinary collaboration, there is much to gain from the multidisciplinary approach. As this paper will argue, combining knowledge and techniques from the two communities can yield results that neither community alone could achieve.

	LINGUISTS	ENGINEERS
uninteresting	provides no explanation of cognitive processes.	provides no useful applications.
ad hoc	without theoretical motivation.	must be provided by hand.
ungeneralizable	"techniques that help you climb a tree may not help you get to the moon."	expense of knowledge engineering prohibits assessing new or more complex domains.

Automatic Speech Understanding, General Issues

Activity and results in automatic speech understanding have increased in recent years, largely because of the "arranged marriage" by a DARPA (Defense Advanced Research Projects Agency; now ARPA) program manager of two previously independent programs: speech recognition and natural language understanding. The speech recognition program was focussed on the automatic transcription of speech, while the natural language understanding program was focussed on interpreting the meanings of typed input. While there are psychological and scientific reasons to integrate these two areas, there are technical and cultural reasons for their past and present degree of separation.

In the ARPA speech understanding program of the 1970s (see, e.g., Klatt's 1977 survey), artificial intelligence (AI) was a relatively new field with much promise.

Expert systems based on speech and language knowledge were developed by separating knowledge sources along traditional linguistic divisions: e.g., acoustic phonetics, phonology, morphology, lexical access, syntax, semantics, discourse. Each module had well-defined inputs and evaluation criteria. A key weakness of the approach, however, turned out to be the number of modules and the decision-making process. Each module may have done fairly well as assessed independently, but when each module was forced to make irrevocable decisions without interaction with other modules, errors could only propagate; a seven-stage serial process in which each module is 90% accurate has an overall accuracy of less than 50%. As statistical pattern matching techniques appeared to perform much better with much less research investment than did the knowledge-based approaches, the funding focus shifted.

Although knowledge-based approaches and statistically based approaches were espoused by people in already largely nonoverlapping communities, these historical events led to larger separation. The "knowledge" proponents argued that the statistical methods only worked accidentally in the short term, on this limited task. The "statistics" proponents assumed that they had been misled by the promise of AI and that they had little to learn from the "knowledge" group. The two communities seemed to come to an agreement that they were working on different problems and had little to say to or ask of each other.

The field of natural language understanding came to be populated largely by "computational linguists" trained in AI techniques in computer science departments for the most part, while speech recognition came to be populated mostly by engineers. The methods, goals, evaluation criteria, background assumptions, and cultures of these two communities are quite different. In fact, a basic disagreement persists over what counts as science. The "knowledge" side values argumentation style, ideas, and long-term research. In their view, the "statistical" side is not scientific because it represents mere engineering "tweaks." The "statistical" side values measurable results. In their view, the "knowledge" side is not scientific because it does not measure results (insofar as the long term tends never to come).

In fact, however, the cause of the differences in performance between the two approaches during the 70s is likely to be an insight of value to both sides: making hard (irrevocable) decisions early, i.e., before considering more knowledge sources, can degrade performance severely. It happened that statistical models provided a mechanism that enabled delayed decision making, and subsequent hardware and algorithmic developments enabled for the consideration of increasingly larger sets of hypotheses. The remainder of this paper will survey the techniques used in combining linguistic with statistical analyses, the issues of interest, and recent results in speech recognition, natural language understanding, and their integration.

Speech Recognition

For several years, the best performing speech recognition systems have been based on statistical pattern matching techniques (Pallett et al. 1990, Pallett 1991, Pallett et al. 1992, Pallett et al. 1993, Pallett et al. 1994). These models are constrained to various degrees by "knowledge" about speech and language (e.g., the topology of the models, the units modeled, the pronunciations modeled, etc.). The most commonly used method is probably hidden Markov models (HMM) (see, e.g., Bahl et al. 1983, Rabiner 1989, Picone 1990). There is also much current work using other pattern matching techniques (see, e.g., Ostendorf and Roukos 1989, Zue et al. 1992), including neural network-based approaches (see e.g., Hampshire and Weibel 1990) and hybrid HMM/neural network approaches (see e.g., Abrash et al. 1994). One can think of the models as representing the linguistic or other knowledge (e.g., what are the units? how are they determined? what aspects need to be represented explicitly? what features will represent the data?). The parameters can then be estimated automatically, given the data and the constraints embedded in the model, to model our "ignorance" — those aspects we can't or don't want to model explicitly.

A Markov model represents the probabilities of sequences of units, e.g., words or sounds. The "hidden" Markov model, in addition, models the uncertainty of the current "state". By analogy with speech production, and using phones as states, the mechanism can be thought of as modeling two probabilities associated with each phone: the probability of the acoustics given the phone (to model the variability in the realization of phones), and the probability of transition to another phone given the current phone. Though some HMMs are used this way, most systems use states that are smaller than a phone (e.g., first, middle, and last part of a phone). Such models have more parameters, and hence can provide greater detail, which can be used to better model duration and context effects. Adding skips and loops to the states can model the temporal variability of the realization of phones. Given the model, parameters are estimated automatically from a corpus of data. Thus models can be "tuned" to a particular (representative) sample.

In this example, the words, phones, and states chosen for the model are units that can be manipulated symbolically (for example, in a set of rules for generating pronunciations based on base forms provided by a dictionary), and they may be theory-driven, data-driven, or some combination. The limited use of linguistic theory in deriving these symbolic components is probably largely a function of the cultural difficulties discussed above. Bridging the cultural gap will require closer collaboration of the two communities: linguists formulating theory in ways that these methods can use, engineers formulating methods that can better capture these knowledge sources.

Typically the HMMs for phones are conditioned on the context (of the previous and/or following phone or phone-class or word, for example). Knowledge of speech and linguistics can provide the choice of the phone set (e.g., Are flap and stop realizations of /t/ modeled as one unit or two?), the topology of the models (e.g., How many states per phone model? Do longer phones have more states, or are loops on certain states sufficient?), and the number of units (e.g., Will phones be modeled as a function of their context? If so, what is the nature of the conditioning context?).

Individual HMMs for phones can be concatenated to model words. Linguistic knowledge, perhaps in the form of a dictionary, determines the sequence of phones that make up a word. Linguistic knowledge in the form of phonological rules can be used to model possible variations in pronunciation, such as the flap or stop realization of /t/. For computational efficiency (at the expense of storage), additional pronunciations can be added to the dictionary. This solution is not ideal for the linguist, since different pronunciations of the same word are treated as totally independent even though they may share all but one or two phones. It is also not an ideal solution for the engineer, since there may be many more parameters to estimate, and recognition accuracy may be lost depending on the implementation, since words with more pronunciations may be disfavored relative to those with few pronunciations. The work of Cohen (e.g., Cohen et al. 1987, Cohen 1989) and others (see, e.g., Withgott and Chen 1993) attempts to address some of these issues, but solutions are not simple and significant performance gains have been hard to come by. Perhaps as researchers are forced to deal with more spontaneous speech effects (as opposed to read speech and highly planned "push-to-talk" speech), these difficult issues will force engineers and linguists to work together to find better solutions.

Linguistic knowledge may also be used to model the effects of lexical stress on vowels. The number of models for vowels could be simply doubled: each vowel has a model representing lexical stress and no lexical stress. Although this captures the linguistic knowledge that lexical stress has an acoustic effect on a vowel, it is not linguistically elegant in that it models the effects of lexical stress on each vowel as independent. The linguist dislikes the solution because it does not capture a generalization about the effect of lexical stress across vowels. The engineer dislikes it because it doubles the number of vowels to model and may not be worth slight gains in performance. This is another example in which the structure of the model constrains how people think about a problem, and in which linguistic and engineering expertise are needed to arrive at a solution.

Modeling utterances the way words were modeled, i.e., a dictionary of all "possibilities," would be even more impossible than it is for words. A list of all possible utterances and their component words would quickly exhaust our resources. However, for limited applications, this solution can be used to simulate continuous speech (i.e., simply model "words" that are very long — namely, long enough to be utterances). A simple (and generally much cheaper) approach is to model all the words in parallel and add a loop from the end to the beginning, where one of the "words" is the "end-of-sentence" word so that the sentences are not infinitely long. Of course, this simple model has the disadvantage of assuming that the ends of all words are equivalent (the same state). This model assumes that at each point in an utterance, all words are equally likely, which we know is not true of any language. Sequences of words can be modeled by concatenating the word models and estimating the probability of different word sequences. The Markov model (minus the hidden part) estimates the likelihoods of words given the previous word (or N words), based on a training corpus of sentence transcriptions. In this example, little linguistic knowledge is used except the intuition that some sequences are more likely than others. That intuition is difficult to call "linguistic" insofar as many linguists work exclusively with grammars in which sentences are either grammatical or not. Though there may be some recognition of doubtful cases, grammaticality is typically a binary decision. Statistical modeling of linguistically relevant relationships is a growing area of interest, though there remain significant technical and cultural challenges.

A survey of recent speech recognition papers reveals the engineering bias (and relative lack of linguistic motivation) in much recent work. Although the major issues facing speech recognition research today include both symbolic and statistical aspects, effective use of both aspects will require increased bridging of the cultural gaps between linguists and engineers. Examples of current speech recognition issues include:

- *Features.* The raw speech waveform needs to be digitized for analysis, and if it is simply sampled in time and amplitude, there is far too much data to handle directly; some feature extraction is needed. The output of feature extraction (sometimes called the "front-end") is the input to the recognition search (which can be based on HMMs, neural nets, or on some other technique). The most common features extracted are cepstral coefficients (derived from a spectral analysis), and derivatives of these coefficients. Sporadically, and especially with reference to noise robustness, there has been interest in improving front-ends and in auditory modeling. Little work has been done since the 70s, however, in modeling linguistically motivated features (e.g., high, low, front, back). Explicit detection of these features has proved challenging. However, a representation of phones in terms of a smaller set of features would have several advantages: fewer parameters could be better estimated, given a fixed corpus; phones that are rare or unseen in the corpus could be estimated on the basis of the more frequently occurring features that compose them; and since features tend to change more slowly than phones, it is possible that sampling in time could be less frequent.

- *Distributions.* In the HMM formulation, the state output distributions have been a topic of research interest recently. The issue under debate has been the use of discrete state distributions, continuous distributions, or (currently the most popular) a mixture of Gaussian distributions (see, e.g., Digalakis and Murveit 1994). Little role for linguistic knowledge is apparent in this work.
- *Model Inventory.* An area of growing interest recently is the choice of units to model. Many systems simply model phones, or phones conditioned on the surrounding phonetic context. Other systems, however, claim improved performance through the selection of units or combination of units determined automatically or semiautomatically (see, e.g., Bahl et al. 1991).
- *Language Modeling.* Any method that can be used to constrain the sequences of occurring words can be thought of as a language model. Traditional grammars are, for example, a type of language model, but so are the Markov models (N-grams) that model only local constraints. As mentioned above, it is a ripe area of research. The goal is to develop language models that can be created easily and can improve speech recognition performance by modeling linguistically motivated attributes (for example, number agreement of subject and verb; or co-occurrences of adjectives with nouns, which may be an arbitrary number of words away from each other) rather than the accidents of word sequences typically estimated by Markov models.
- *Adaptation.* The first speech recognition systems tended to be speaker dependent (before using a system, a person had first to read a list of words or sentences). In recent years, the trend has been toward speaker independence. Speaker-independent systems can work reasonably well for a variety of talkers, but the broader coverage of talker types, dialect types, and so on, the more fuzzy the models. The future is likely to be in speaker adaptation: the system begins as a speaker-independent system and gradually adapts to the characteristics of a new speaker. This approach is not unlike linguistic experience in which new dialects may be difficult to understand at first. In a foreign language it may be easier to observe adaptation to individual speakers.
- *Search.* Given the acoustic models, the language models, and the input speech, the role of the recognizer is to search through all possible hypotheses and find the best (most likely) string of words. As the acoustic and language models become more detailed they become larger, and this can be an enormous task, even with increasing computational power. Significant effort has been spent on managing this search: depth-first vs. breadth first, beam search (which prunes hypotheses if they are enough below the best-scoring hypothesis), and, recently, various schemes for making multiple passes using coarser models at first to narrow the search and progressively more detailed models to further narrow the pruned search space (see, e.g., Murveit et al. 1993, Nguyen et al. 1993).
- *Robustness.* Robustness is a key area of research. Systems can be developed that function well in narrow contexts, but to be useful in a wide range of applications, they need to be robust to the variability that occurs in speech communication: variability due to differences in talkers, speech styles, microphone and noise conditions, dialect, and language. This is an area in which the forcing function of hard problems may help to bridge the cultural gaps, as engineers realize that narrowing the solution decreases robustness and requires the more general solution sought by linguists and speech scientists.
- *Portability.* Portability is another key area of research. Creating a demonstration in a limited domain may give the feel of accomplishment, but science (and applications) demand generalizability and reproducibility. We cannot imitate the range of human capability with speech recognition systems, but we can create useful applications in limited domains. The amount of work we have done for one task that can be reused in another task is a measure of how much we have learned about speech generally. (As will be argued in a later section, linguists and speech scientists also need to assess the portability of their knowledge.)
- *Scalability.* In an environment in which computer power is rapidly changing (increasing power on platforms of the same size, and decreased power on ever smaller platforms), another key issue is scalability: the capability of using increased memory and computational power for faster, more accurate recognition on the one hand, and the capability of gracefully degrading on platforms with less memory and computational power. Although linguistic and speech knowledge, as suggested in the examples above, can help form more efficient representations, scalability is not really a linguistic issue.

The gap between speech scientists and speech recognition engineers has meant that some aspects of speech have had to be discovered independently. Many cognitive models appear to be more continuous than they used to be, and are looking a bit more like the statistical models than was previously true. The gap between the two areas has meant that many speech researchers have not been able to take advantage of statistical tools that could help them advance their knowledge. It has also meant that advances in speech research and models of cognition have not been able to affect automatic speech recognition. For example, the notion of a prototype and distance from a prototype (see, e.g., Massaro 1987, Kuhl 1990) which seems to explain much data from speech perception (and other areas of perception), is not well modeled in the current speech recognition frameworks. A person who has not been well understood tends to change the speech style so as to be better understood. This may involve speaking more loudly or more clearly, changing the phrasing, or perhaps even leaving pauses between words. These changes may help in human-human communication, but in typical human-machine interactions, they result in forms that are even more difficult for the machine to interpret. The concept of a prototype in machine recognition could lead to more robust recognition technology.

That is, the maximum-likelihood approaches common in speech recognition miss a crucial aspect of language: the role of contrast. A given linguistic entity (e.g., phone) is characterized not just by what it is, but also by what it is not, i.e., the system of contrast in which it is involved. Thus, hyperarticulation may aid communication over noisy phone lines for humans, but may decrease the performance of recognizers trained on a corpus in which this style of speech is rare or missing. The results can be disastrous for applications, since when a recognizer misrecognizes, a common reaction is to hyperarticulate (Shriberg *et al* 1992). Discriminative systems, such as neural network formulations, have an advantage over maximum-likelihood approaches in this respect, though it is an area in which linguists and speech perception experts could play a larger role.

Although things are changing rapidly, and many factors will affect just how well a system will perform, examining recent benchmark evaluations can give an idea of the relative difficulty of various aspects of speech (see e.g., Pallett et al. 1994). These areas could be those in which increased linguistic knowledge could improve performance. For example, the variance across the talkers used in the test set was greater than the variance across the systems tested. Further, the various systems tested had the highest error rates for the same three talkers who were the fastest talkers in the set. These observations could be taken as evidence that variability in pronunciation, at least insofar as fast speech is concerned, may not currently be well modeled. Further evidence of the need for better modeling of the pronunciation variation observed in spontaneous speech arises from the degradation in recognition accuracy observed in moving from read speech or carefully planned speech to normal, conversational speech.

Natural Language Understanding

The field of natural language (NL) understanding has been traditionally populated by computational linguists, trained in artificial intelligence, largely in computer science departments. The approaches have traditionally been based in symbolic logic, using expert-systems techniques typically involving large sets of hand-crafted rules. The arranged "marriage" with speech recognition has resulted in a great increase in the use of statistical methods for automatically creating natural language components, or for automatically training their parameters. Since the first joint meeting of the speech and natural language communities in 1989, the number of papers and the range of topics addressed using statistical methods have steadily increased. At the most recent meeting (March 1994), the category of statistical language modeling and methods received the most abstracts and was one of the most popular sessions.

The issues of concern in natural language research are largely determined by the interests of those doing the research, and at present they tend to be computational linguists. However, as argued above, there is a growing tendency to combine knowledge-based with statistical/engineering approaches. Based on recent papers, topics of major concern include:

- *Lexicon.* Although speech recognition components usually use a lexicon, lexical tools in natural language are more complex than lists of words and pronunciations. Different formalisms store different types and formats of information, including, for example, morphological derivations, part-of-speech information, and syntactic and semantic constraints on combinations with other words. There is little evidence, however, in most of these representations that some structures are more likely than others.
- *Grammar.* A grammar is typically a set of rules devised by observation of occurring patterns in a language or sublanguage. Typically, grammars either accept a sentence or reject it, although grammars that degrade more gracefully in the face of spontaneous speech and recognition errors are being developed (see, e.g., Hindle 1992). Another issue of relevance is the development of grammars that can be used either for analysis (parsing) or for generation. This should become increasingly important as machines play a more active role in human-machine collaboration.
- *Robustness.* Robustness has been a major issue in recent years in natural language. The traditional computational linguistic approach of covering a set of linguistically interesting examples was put to a severe test in the attempt to cover, in a limited domain, the set of utterances produced by people engaged in problem-solving tasks. Several new sources of complexity were introduced: the move to an empirically based approach (covering a seemingly endless number of "simple" things became more important than covering the "interesting," but more rare, complex phenomena), the separation of test and training materials (adding rules to cover phenomena observed in the training corpus may or may not affect coverage on an independent test corpus), the nature of spontaneous speech (which has a different, and perhaps more creative, structure than written language, previously the focus of much NL work), and recovery

from errors that can occur in recognition.

- *Parsing.* The goal of parsing is to retrieve or assign a structure (based on the grammar used) to a string of words for use by a later stage of processing. Typically, parsers have worked deterministically on a single string of input. When parsers were faced with typed input, aside from the occasional typo, the intended words were not in doubt (though their parts of speech or syntactic role might have been). When speech is the input, however, speaker disfluencies, novel syntactic constructions and recognition errors pose serious difficulties for traditional parsers.
- *Interpretation.* Interpretation may or may not be separated from parsing. Typically, however, parsing is faster than interpretation and narrows the field considerably for the interpretation stage. Interpretation is the stage at which a representation of meaning is constructed. Of course, this representation is not of much use without a "back-end" that can use the representation to perform an appropriate response, e.g., retrieve a set of data from a database, ask for more information, etc. This stage is typically purely symbolic, though likelihoods or scores of plausibility may be used.
- *Portability.* Portability has been less of a research area in NL than in speech recognition, largely because many of the methods used are so costly (data collection for speech recognition can be costly as well, but it may be argued that it can be done by nonexperts). The portability issue can be expected to grow in importance in NL work. Automatic acquisition and automatic tuning of parameters are already growing areas of research, representing the impact of cross-disciplinary fertilization (see, e.g., the recent ARPA Human Language Technology Workshop proceedings).
- *Scalability.* As for speech research, scalability becomes increasingly an issue as the technology becomes appropriate for technology transfer. Even for demonstrations of feasibility, it can be important to develop algorithms that run quickly enough on a platform small enough to be widely available.

The combining of traditionally linguistic or AI approaches with statistical modeling techniques, as already mentioned, is more or less involved in all the issues just outlined. Although difficult, such cross-disciplinary work still holds much promise for future advances. Recent trends in ARPA proceedings papers indicate that new uses of statistics in NL areas far outnumber new uses of linguistics in speech recognition. Perhaps the difficulties posed by conversational spontaneous speech will cause engineers to take another look at linguistics.

Results in natural language understanding have been more resistant to quantification than those in speech recognition (where there is fairly good agreement on the string of words produced). What does it mean to have understood properly? Can there be more than one way to understand properly? In the ARPA community, these hard questions have been postponed somewhat by agreeing to evaluate on the answer returned from a database. Trained annotators examine the string of words (NL input) and use a database extraction tool to extract the minimum and maximum accepted set of tuples from the evaluation database. A "comparator" then automatically determines whether a given answer is within the minimum and maximum allowed.

The community is not, however, content with the current expense and limitations of the evaluation method described above, and is investing significant resources in finding a better solution. Key to much of the debate is the cultural gap: engineers are uncomfortable with evaluation measures that cannot be automated (forgetting the role of the annotator in the current process); and linguists are uncomfortable with evaluations that are not diagnostic; and, of course, neither side wants significant resources to go to evaluation that would otherwise go to research.

Integration of Speech Recognition and Natural Language Understanding

The integration of the two technologies outlined in the previous sections seems to be a natural connection. Nonetheless, the two communities were distinct enough that, except for the funding impetus, the coupling might not have happened. Many researchers in both communities would agree however, that the integration effort has been good for both. To natural language understanding, speech recognition can bring prosodic information, information important for syntax and semantics but not well represented in text. NL can bring to speech recognition several knowledge sources (e.g., syntax and semantics) not previously used (N-grams model only local constraints, and largely ignore systematic constraints such as number agreement). For both, the integration affords the possibility of many more applications than could otherwise be envisioned, and the acquisition of new techniques and knowledge bases not previously represented.

One of the main lessons of the ARPA speech understanding project of the 1970s was that considering all knowledge sources before making hard decisions was a big win. In speech recognition, tighter integration has consistently led to improved performance. However, technical and cultural differences inhibit such tight integration.

Technically, as coverage increased, language models tended to grow and required either increasingly large amounts of storage (possibly infinite, if all rules were to be compiled, and some rules had infinite loops), or increasingly large amounts of computation (if an interpretive approach was envisioned). Thus, the prospect of NL guiding the recognition search became nearly hopeless; rather, it appeared that speech output was needed to guide the NL search if the task was to be done quickly enough. Further, increased coverage also meant that NL grammars provided less constraint, and constraint of a type that made it difficult to prune recognition hypotheses early in the speech stream. Although long-distance constraints like this were desired, the solutions resulted in too many active speech hypotheses to be manageable.

The cultural difficulties have already been outlined, and as technical difficulties settled in, there was a tendency to take the easy way out and settle into the arranged marriage with separate bedrooms: a strictly serial approach defined the turf, and each side focussed on improving its own technologies, though with some exposure to the techniques and culture of the other side.

In this climate of compromise, the N-best integration approach became popular. In this approach, the connection between the two components is strictly serial, but the hard-decisions-early issue is softened by sending not just the best hypothesis from speech recognition, but the N-best (where N may be on the order of 10 to 100). The NL component can then score this set for grammaticality (where in some cases the "score" is just a 1 for "grammatical" or a 0 for "not grammatical"), and combine the acoustic with the grammar score to determine the best-scoring hypothesis. This approach is computationally tractable, and accomodates great modularity of design (different speech and NL modules can be swapped in and out). Examples of N-best interfaces include: Veilleux and Ostendorf 1993, Rayner 1994.

The limitations in the N-best integration are related to the modules: (1) if the speech module is not very accurate, N may have to be very large to ensure that the correct string is included; this becomes more of a problem as vocabulary size, noise, or other parameters that degrade performance increase; (2) an NL component that provides only a score of 1 or 0 is limited in its ability to take advantage of the N-best outputs, particularly for large N, since many of the N-best may be grammatical, and some will be more grammatical than others. A strategy to combat this problem is a lattice-based interface (Murveit et al. 1993). A lattice of speech hypotheses can compactly include a very large N and greatly improve computational efficiency, especially for parsers that can parse lattices.

The integration of speech recognition and NL is concerned with many of the same issues that each of the components face: robustness, portability, speed, and size. This section has so far outlined some issues that arise in designing the architecture for combining the speech and NL. However, the integration gives rise to some new areas as well: how can an NL component deal with spontaneous speech effects such as false starts and repairs and how can a speech component send information to help the NL component (see e.g., Bear et al. 1992, Shriberg et al. 1992), how can techniques from the two component areas be effectively combined, and how can prosodic information be effectively communicated between the two components (see, e.g., Price and Ostendorf 1994 for survey).

Speech understanding in some sense works as well as its two components if they are serially connected. However, performance can be maximized if the two components take into account the strengths and weaknesses of each other. In the ARPA benchmarks, if we compare on the same testset, we find that the best speech recognition results provide a completely correct transcription of the utterance less often than the speech understanding results (speech recognition plus NL) provide a correct answer. This condition arises because many of the speech errors do not affect the correctness of the answer (e.g., "flight" vs. "flights", "a" vs. "the"), and because the understanding components have become more robust to speech recognition and speaker errors, false starts, and neologisms. This situation is not unlike human speech understanding (particularly apparent over the telephone or in a language that is not your native language), when you can better make out the sense of what is meant than give an exact transcription of what was said.

The results compiled in the ARPA benchmark papers document the state of the art. However, it is not at all clear that these are the right measures, or at least the only right measures. In separate experiments, we have tried to correlate the "correctness" of the system's answer with subsequent user behavior. We have found many factors that affect user behavior in predictable ways, but the correctness or incorrectness of the answer seems to have little effect. While it might be that we just have not yet found the right measure, there are several reasons that this correlation may be difficult to obtain. For example, the user may or may not notice that the answer is incorrect; the system's answer may be incorrect but provide a superset of the information requested so that the user can continue without measurable interruption; or the system's answer could be correct but look incorrect to the user, either because of user error, or because a mistake in understanding happens to result in a correct answer (as frequently happens when a day of the week is misrecognized, since information on different days is the same). In particular, we know that performance is affected by factors such as vocabulary size, task complexity, noise conditions, but we do not know how to generalize results from a particular benchmark condition to those in which all these parameters may differ. Complementary measures to "correct answer" include: user satisfaction, time to complete standard tasks, user preference, and (perhaps the bottom line) units sold. The speech understanding community is quite active in refining evaluation measures and in developing new ones (see e.g., Price et al. 1992, Hirschman et al. 1993, Dahl et al. 1994). because the evaluation measures guide the research directions, it is important to choose the right measures,

Discussion and Summary

This audience should not require motivation of symbolic and traditionally linguistic knowledge sources, although some may have hesitation about statistically based engineering approaches. This paper is an attempt to help bridge the gap between the largely symbolic and the largely statistical approaches. Although statistical models are far from the only tool for investigating speech and language, as argued in Price and Ostendorf (1994) they do provide several important features: they can be trained automatically, they can provide a systematic way to combine multiple knowledge sources, they can express the more continuous properties of speech and language, they make it easier to deal with large corpora, they provide a means for assessing incomplete knowledge, and they provide a means for acquiring knowledge about speech and language.

The ability to consider large corpora is a particularly valuable attribute. Large corpora are the only place rare phenomena will be found in sufficient number to be studied adequately. Large corpora offer the possibility of mediating two competing trends in speech and language research: "ecological validity" (which acknowledges that any change in conditions can affect the data, and therefore limits the data to speech and language occurring in conditions as natural as possible) and "speech science" (which acknowledges that any change in conditions can affect the data, and therefore limits the data to speech and language occurring in strictly controlled environments such as sound-proof booths, and read speech.). Both sides start with the same premise and choose opposite approaches. Because language is so rich and variable, there will continue to be a need for both approaches. However, large corpora offer a data point somewhat in between: if the variable of interest recurs frequently enough, large corpora can provide enough naturally occurring instances to "wash out" the effects of the various environments in which it occurs. Without automatic methods, many of them involving statistics, large corpora would be impossible to analyze.

As argued in Price and Ostendorf (1994), the increasingly popular classification and regression trees, or decision trees (see, e.g., Breiman et al. 1984) appear to be a particularly useful tool in bridging the cultural and technical gap in question. In this formalism, the speech researcher or linguist can input the types of information that are known to affect variability (duration of a phone, for example), and based on a corpus of data in which these parameters are observed, the resulting tree can show how much of the variability is accounted for by each source of information (for example, voicing of following consonant, compared to existence of following silence). Examples of the use of this tool are numerous: e.g., Hirschberg 1993, Ostendorf and Veilleux 1993, Wang and Hirschberg 1991, and Withgott and Chen 1993.

Of course, the biggest disadvantage of many of the existing statistical and other engineering models is cultural discomfort. New techniques structure the way one thinks about problems, and this can be uncomfortable and even threatening. However, the advantages offered by multidisciplinary approaches are large. Obviously, the gap can be bridged by becoming fluent in the new techniques, but this is increasingly difficult as the challenges of keeping up with existing fields increase. The gap can also be bridged by collaboration with others who are already fluent in the techniques, and by encouraging students to learn more about the techniques.

In sum, combining statistical with linguistic models has led to important gains in speech recognition and speech understanding, and to more powerful tools for acquiring further knowledge. Fuller understanding will require knowledge that spans all linguistic levels, from acoustics through semantics and pragmatics/discourse. Few people are trained in all these areas. Fewer still have training in statistical methods. Therefore, in the near term, multidisciplinary collaborations will be essential for rapid progress.

ACKNOWLEDGMENT. I thank Mari Ostendorf for her useful comments on the manuscript. I gratefully acknowledge the support of ARPA/ONR Contract ONR N00014-90-C-0085, and ARPA/NSF funding through NSF Grant IRI-8905249. The opinions expressed are the author's and not necessarily those of the funding agencies.

References

[1] Abrash, V., M. Cohen, H. Franco, and I. Arima (1994) "Incorporating Linguistic Features in a Hybrid HMM/MLP Speech Recognizer," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* 62.8.1-4.

[2] Bahl, L., Jelinek, F. and Mercer, R. L. (1983). "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence PAMI-5,* 2, 179–190.

[3] Bahl, L., P. de Souza, P. Gopalakrishnan, D. Nahamoo, and M. Picheny (1991) "Context Dependent Modeling of Phones in Continuous Speech using Decision Trees," *Proceedings of the DARPA Speech and Natural Language Workshop,* pp. 264-269.

[4] Bear, J., J. Dowding, and E. Shriberg (1992) "Integrating Multiple Knowledge Sources for Detection and Correction of Repairs in Human-Computer Dialog," *Proc. of the Annual Meeting of the Association for Computational Linguistics,* pp. 56-63. Also published as SRI Technical Note 518.

[5] Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984) *Classification and Regression Trees,* Wadsworth and Brooks/Cole Advanced Books and Software, Monterey, CA.

[6] Cohen, M., G. Baldwin, J. Bernstein, H. Murveit and M. Weintraub (1987) "Studies for an Adaptive Recognition Lexicon," *Proceedings of the DARPA Speech and Natural Language Workshop,* pp. 49-55.

[7] Cohen, M. (1989) '*Phonological Structures for Speech Recognition*, Department of Computer Science, University of California at Berkeley, Ph.D. Dissertation, University of Michigan Microfilms.

[8] Dahl, D., M. Bates, M. Brown, W. Fisher, K. Hunicke-Smith, D. Pallett, C. Pao, A. Rudnicky, and E. Shriberg (1994) *Proceedings of the ARPA Human Language Technology Workshop,* to appear.

[9] Digalakis, V., and H. Murveit (1994) "An Algorithm for Optimizing the Degree of Tying in a Large Vocabulary Hidden Markov Model Based Speech Recognizer," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* 54.2.1-4.

[10] Hampshire, J. and A. Weibel (1990) "Connectionist Architectures for Multi-Speaker Phoneme Recognition," in D. Rouretzky (ed.) *Advances in Neural Information Processing Systems 2*, Morgan Kaufman.

[11] Hindle, D. (1992) "An Analogical Parser for Restricted Domains," *Proceedings of the ARPA Human Language Technology Workshop,* pp. 150-154.

[12] Hirschberg, J. (1993) "Pitch Accent in Context: Predicting Prominence from Text," *Artificial Intelligence,* Vol. 63, No. 1-2, pp. 305-340.

[13] Hirschman, L., L. Bates, D. Dahl, W. Fisher, J. Garofolo, D. Pallett, K. Hunicke-Smith, P. Price, A. Rudnicky, and E. Tzoukermann (1993) "Multi-Site Data Collection and Evaluation in Spoken Language Understanding," *Proceedings of the ARPA Human Language Technology Workshop,* pp. 19-24.

[14] Klatt, D. (1977) "Review of the ARPA Speech Understanding Project" *J. Acoust. Soc. Amer. 62*, no. 6, pp. 1345–1366.

[15] Kuhl, P. (1990) "Towards a New Theory of the Development of Speech Perception," *Proc. Int. Conf. on Spoken Language Processing 2*, pp. 745-748.

[16] Massaro, D. (1987) *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry,* Hillsdale, NJ, Lawrence Erlbaum Associates.

[17] Murveit, H., J. Butzberger, V. Digalakis, and M. Weintraub (1993) "Large Vocabulary Dictation using SRI's DECIPHER Speech Recognition System: Progressive Search Techniques," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* pp. II-319-322.

[18] Nguyen, L., R. Schwartz, F. Kubala, and P. Placeway (1993) "Search Algorithms for Software-Only Real-Time Recognition with Very Large Vocabularies," *Proceedings of the ARPA Human Language Technology Workshop*, pp. 91-95.

[19] Ostendorf, M. and S. Roukos (1989) "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, December, pp. 1857-1869.

[20] Ostendorf, M. and N. Veilleux (1993) "A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary Location," *Computational Linguistics.*

[21] Pallett, D. (1991) "DARPA Resource Management and ATIS Benchmark Test Poster Session," *Proc. Speech and Natural Language Workshop*, Morgan Kaufman, pp. 49–58.

[22] Pallett, D., N. Dahlgren, J. Fiscus, W. Fisher, J. Garofolo and B. Tjaden (1992) "DARPA February 1992 ATIS Benchmark Test Results," *Proc. Speech and Natural Language Workshop*, Morgan Kaufman, pp. 15–27.

[23] Pallett, D., J. Fiscus, W. Fisher, J. Garofolo, B. Lund, and M. Prysbocki (1994) "1993 Benchmark Tests for the ARPA Spoken Language Program," *Proc. Human Language Technology Workshop*, Morgan Kaufman, to appear.

[24] Pallett, D., J. Fiscus, W. Fisher, and J. Garofolo (1993) "Benchmark Tests for the DARPA Spoken Language Program," *Proc. Human Language Technology Workshop*, Morgan Kaufman, pp. 7–18.

[25] Pallett, D., W. Fisher, J. Fiscus and J. Garofolo (1990) "DARPA ATIS Test Results," *Proc. Speech and Natural Language Workshop*, Morgan Kaufman, pp. 114–121.

[26] Picone, J. (1990) "Continuous Speech Recognition Using Hidden Markov Models," *IEEE ASSP Magazine*, pp. 26-41.

[27] Price, P., L. Hirschman, E. Shriberg, E. Wade (1992) "Subject-Based Evaluation Measures for Interactive Spoken Language Systems, " *Proceedings of the DARPA Speech and Natural Language Workshop*, pp. 34-38.

[28] Price, P. and M. Ostendorf (1994) "Combining Linguistic with Statistical Methods in Modeling Prosody," in J. L. Morgan and K. Demuth (Eds.), *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

[29] Rabiner, L. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition," *IEEE Proceedings 77,* 2, 257–286.

[30] Rayner, M., D. Carter, V. Digalakis, and P. Price (1994) "Combining Knowledge Sources to Reorder N-Best Speech Hypothesis Lists," *Proceedings of the ARPA Human Language Technology Workshop,* (to appear).

[31] Shriberg, E., J. Bear, and J. Dowding (1992) "Automatic Detection and Correction of Repairs in Human-Computer Dialog," *Proceedings of the DARPA Speech and Natural Language Workshop,* pp. 419-424.

[32] Shriberg, E., E. Wade, P. Price (1992) "Human-Machine Problem Solving Using Spoken Language Systems (SLS): Factors Affecting Performance and User Satisfaction," *Proceedings of the DARPA Speech and Natural Language Workshop,* pp. 49-54.

[33] Veilleux, N. and M. Ostendorf (1993) "Probabilistic Parse Scoring with Prosodic Information," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* pp. II51-55.

[34] Wang, M. and J. Hirschberg (1991) "Predicting Intonational Boundaries Automatically from Text: The ATIS Domain," *Proceedings of the DARPA Speech and Natural Language Workshop,* pp. 378-383.

[35] Withgott, M. and F. Chen (1993) *Computational Models of American Speech,* CSLI Lecture Notes Number 32.

[36] Zue, V., J. Glass, J. Goddeau, D. Goodine, L. Hirschman, H. Leung, M. Phillips, J. Polifroni, and S. Seneff (1992) "The MIT ATIS System: February 1992 Progress Report," *Proceedings of the ARPA Human Language Technology Workshop,* pp. 84–88.