FROM SIGNALS TO SYMBOLS TO MEANING: ON MACHINE UNDERSTANDING OF SPOKEN LANGUAGE[1]

Victor W. Zue

Laboratory for Computer Science, Massachusetts Institute of Technology
Cambridge, Massachusetts 02139, U.S.A.

ABSTRACT

This paper starts with a brief overview of advances in the development of speech recognition systems, with particular emphasis on the past decade. It then moves on to make two points. First, successful development of speech recognition systems will depend on our ability to understand human communication through spoken language, to capture the essential features of the process in appropriate models, and to develop the necessary computational framework to make use of these models for machine understanding. Second, just as human communication using spoken language is an active process of understanding, we must begin to investigate methods that will combine speech recognition and natural language processing technology to achieve speech understanding. Examples to support these arguments will be provided.

INTRODUCTION

Spoken language is the most natural, flexible, efficient, and economical means of communication among humans. As computers continue to play an increasing role in our lives, it is important that we seriously address the issue of providing a graceful human-machine interface through spoken language. Research in speech coding and synthesis has matured over the past decade to the extent that speech can now be transmitted efficiently and generated with high intelligibility. Spoken input to computers, however, has yet to cross the threshold of practicality. To be sure, the last decade has witnessed dramatic improvement in speech recognition technology. Nevertheless, current speech recognition systems still fall far short of human capabilities of continuous speech recognition with essentially unrestricted vocabulary and speakers, under difficult acoustic conditions.

Why is it so hard to develop computer systems to recognize speech? One of the primary reasons is the variabilities that one finds in spoken language communication. Speech can be produced by many speakers with diverse vocal tract anatomies and sociolinguistic backgrounds. Even for a particular speaker, the characteristics of the signal can vary over a wide range, depending on his or her physiological and psychological states. Many external factors, such as the acoustic environment and the types of microphone can also significantly alter the resulting signal. One may be tempted to dismiss these variabilities as undesirable *noise* imposed on the otherwise invariant signal. In reality, however, the process of encoding linguistic information in spoken language is highly stochastic in nature. Speakers of a language can convey the same underlying message with many choices of words and linguistic constructs. Furthermore, even though the inventory of phonemes for a language is quite small, their acoustic-phonetic realizations depend critically on the context in which they appear, as illustrated in Figure 1. For example, while the initial /t/ in the words "two" and "ten" share some acoustic similarities, there are also significant differences that one can readily observe. The burst release for the first /t/ is lower in frequency than the second, a direct consequence of anticipatory coarticulation caused by the following rounded vowel /u/. By the same token, the acoustic similarities of the three /ɛ/'s in the words "seven," "less," and "ten" are overshadowed by the apparent differences. The second /ɛ/ shows articulatory undershoot due to lateralization, as evidenced by the lowering of its second formant, whereas the last /ɛ/ is heavily nasalized, indicated by the smearing of the first formant. Figure 1 also contains more subtle examples of contextual variations. For example, the spectra for the alveolar strident fricatives in the words "is" and "less" both tilt upwards near the end, but for apparently different reasons. In the first case, the upward tilt is due to the following lateral consonant, and is often accompanied by a brief period of epenthetic silence, followed by the relatively sudden lateral release. In the second case, the upward tilt is due to the following dental fricative, which has a more anterior place of articulation. To be sure, remarkable advances have been made in various disciplines of phonetic science, so that we now have a far better

[1] This research was supported by DARPA under Contract N00014-89-J-1332, monitored through the Office of Naval Research.

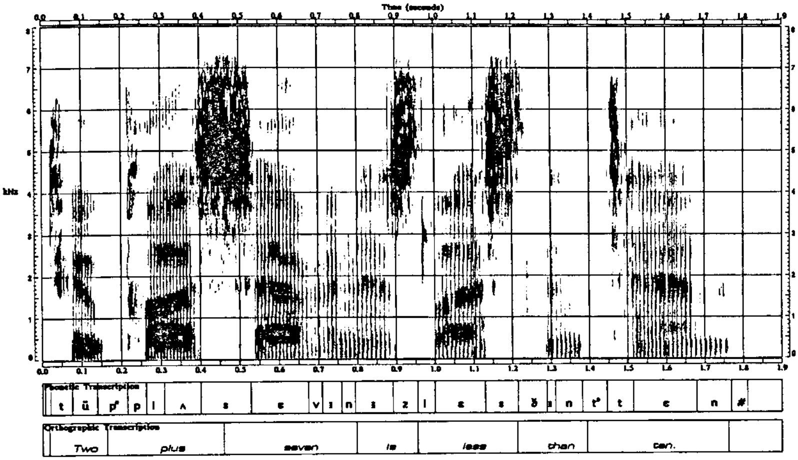


Figure 1: Digital spectrogram of the sentence "Two plus seven is less than ten," spoken by a male talker. Also included are phonetic and orthographic transcriptions that are aligned with important acoustic landmarks in the signal. The spectrogram illustrates some of the acoustic-phonetic variations often found in continuous speech.

understanding of many aspects of this variability than we did a few short decades ago. Nevertheless, researchers in automatic speech recognition have not been able to capitalize on the vast amount of knowledge, primarily because of the lack of a unifying computational framework to make use of it.

I will start this paper with a brief review of the state of the art in speech recognition by machine, with particular emphasis on the past decade. This will be followed by my assessment of the factors contributing to the improvement in systems' performance. I will use the remainder of the paper to make two points. First, successful development of speech recognition systems will depend on our ability to understand human communication through spoken language, to capture the essential features of the process in appropriate models, and to develop the necessary computational framework to make use of these models for machine understanding. Second, just as human communication using spoken language is an active process of understanding, we must begin to investigate methods that will combine speech recognition and natural language processing technology to achieve speech understanding. Indeed, many of the applications of human/machine interface through spoken language require systems possessing the capability of solving a problem interactively with a user. To illustrate my points, I will draw liberally from our own experience in developing speech recognition and speech understanding systems. This is done primarily for the sake of familiarity, and not ethnocentricity.

STATE OF THE ART IN SPEECH RECOGNITION

Defining the Parameters

Speech recognition systems can be characterized by many parameters. An isolated-word speech recognition system requires that the speaker pause briefly between words, whereas a continuous speech recognition system does not. Some systems require speaker enrollment; a user must provide samples of his or her speech before using them. Other systems are said to be speaker-independent in that no enrollment is necessary. Some of the other parameters depend on the specific task. Recognition is generally more difficult when vocabularies are large or have many similar sounding words. The language model is the artificial grammar that restricts the combination of words. The simplest language model can be specified as a finite-state network, where the permissble words following each word are given explicitly. More general language models approximating natural language are specified in terms of a context-sensitive grammar. One popular measure of the difficulty of the task, combining the vocabulary size and the language model, is *perplexity*, P, defined as:

$$P = 2^{-\frac{1}{N}\sum_{i=1}^{N} log_2 P(w_i|w_{i-1},...w_1)}$$

where the w_i are the sequence of all words in all sentences, N is the total number of words, and $P(w_i|w_{i-1},...w_1)$ is the probability of the ith word given all preceding words. Perplexity is related to the average number of words allowed at each node in the language model.[2] Finally, there are some external parameters that can affect speech recognition system performance, including the characteristics of the environmental noise, the type and the placement of the microphone, speaker's level of physiological and psychological stress, and variations in speaking rate.

Performance Review

What follows is a snapshot of current performance of some typical systems on a variety of tasks. It is intended to be illustrative, rather than exhaustive. Interested readers are referred to an extensive review by Mariani for additional information [32].

Performance of speech recognition systems is typically described in terms of word error rate, E, defined as:

$$E = (1 - \frac{S + I + D}{N})100\%$$

where N is the total number of words in the test set, S, I, and D are the total number of substitutions, insertions, and deletions, respectively. Note that insertion and deletion are meaningful measures only for *continuous* speech recognition systems.

Low Perplexity Tasks One of the most popular, and potentially most useful task with low perplexity ($P = 11$) is the recognition of digits. For American English, speaker-independent recognition of digit strings spoken continuously and restricted to telephone bandwidth can achieve an error rate of 0.8% when the string length is known. When the string length is unknown, the error rate increases to 1.4% [48]. This represents a significant improvement over the best systems only a decade ago, which had an error rate of 2%, for digits spoken in isolation by known talkers, recorded under high quality conditions [13]. The French isloated digit recognition system developed at CNET performed robust enough to be deployed over public telephone network [14].

Another potentially useful task is the recognition of English alphabets ($P = 26$). Despite the low perplexity, English alphabet recognition is a very challenging task, since many of the letters are acoustically similar. In 1983, Cole reported an error rate of 10.5% on speaker-independent recognition of isolated digits with a system that makes use of acoustic features known to be important for fine phonetic contrast [9]. Staying with the same philosophy but using an artificial neural net classifier, Cole recently achieved a speaker-independent error rate of 4% for isolated letters of the alphabet [11].

Moderate Perplexity Tasks In the eighties, a number of researchers have pursued speech recognition tasks with a vocabulary of a few hundred words and moderate perplexity. One of the best known is the 1,000-word Resource Management (RM) task, in which inquiries can be made on various naval vessels in the Pacific ocean. This task was made popular by the fact that it is the designated task for common evaluation among contractors of the U.S. Defense Advanced Research Projects Agency's Strategic Computing Program. As a result, speech data for system training and testing, as well as evaluation procedures, have been developed and are readily available [37].

The best speaker-dependent results on the RM task were achieved by BBN and MIT Lincoln Laboratory. Using a word-pair language model that constrains the possible words following a given word ($P = 60$), these systems achieved a word error rate of less than 2% on continuously spoken sentences [38]. The BBN BYBLOS system can also operate in a speaker-adaptive mode, in which the system adapts its models and parameters using only 40 sentences from the new speaker. A 45% reduction in word error rate for the new speaker can be realized with rapid system adaptation [24]. The ARM system developed at RSRE in the United Kingdom achieved an error rate of 13.2% on a 497-word task with no language model ($P = 497$) [42]. For comparison, researchers at IBM reported a word error rate of 9% on the 1,000-word Laser Patent task ($P = 24$) only a few short years ago, and it was the best result at that time [2].

Over the past few years, good performance on speaker-independent recognition for moderate perplexity is beginning to emerge, the best known being the SPHINX system developed at Carnegie Mellon University [27]. On the RM task ($P = 60$), SPHINX achieved a speaker-independent word error rate of 4.5% [38].

High Perplexity Tasks High perplexity tasks with a vocabulary of thousands of words are intended primarily for the dictation application. To make the task manageable and performance reasonable, however, the systems are typically speaker-dependent, and require that the speaker pause between words. Researchers at IBM's T. J. Watson Research Center are among the most active and successful in this area. For example, the TANGORA system achieved word error rate of 2.9% and 5.4% on a 5,000-word and a 20,000 word office dictation task [1]. Similar efforts can also be found in Canada and France [28,34]. The INRS 86,000-word system achieved an error rate of 7.2%, whereas researchers at IBM-France reported an erorr rate of 12.7% on their 200,000-word system.

Discussion

The improvement in speech recognition technology over the last decade was brought on by

[2] Perplexity for the English language has been estimated from text to be between 150 and 200.

several factors. First and foremost, there is the coming of age of the utilization of stochastic modelling techniques. The AT&T digit recognition system, the BYBLOS and SPHINX continuous speech recognition systems, as well as all the high perplexity systems mentioned earlier are all based on some form of hidden Markov modelling (HMM). HMM is a doubly stochastic model, in which the generation of the underlying phoneme string and their surface acoustic realizations are *both* represented probabilistically as Markov processes [41, 40]. HMM is powerful in that, with the availability of training data, the parameters of the model can be trained automatically to give optimal performance. While the application of HMM to speech recognition started nearly twenty years ago [21,3], it was not until the past few years that it gained wide acceptance by the research community.

Second, much work has gone into the development of large speech corpora for system development, training, and testing [6,25,37,53,19,5]. Some of these corpora are designed for acoustic phonetic research, while others are highly task specific. These corpora permit researchers to quantify the acoustic cues important for phonetic contrasts and to determine parameters of the recognizers in a statistically meaningful way. The importance of their availability cannot be overstated.

Third, progress has been brought about by the establishment of standards for performance evaluation. Less than a decade ago, researchers trained and tested their systems using locally collected data, and had not been very careful in delineating training and testing sets. As a result, it is very difficult to compare performance across systems, and the system's performance typically degrades when presented with previously unseen data. The recent availability of a large body of data in the public domain, coupled with the specification of evaluation standards [37], has resulted in uniform documentation of test results, thus contributing to greater reliability in monitoring progress.

Finally, advances in computer technology also indirectly influenced our progress. The availability of fast computers with inexpensive mass storage capabilities has enabled many researchers to run many large scale experiments in a short amount of time. This means that the elapsed time between an idea and its implementation and evaluation is greatly reduced.

INCORPORATING SPEECH KNOWLEDGE

Successful systems developed over the last decade are very different from their predecessors. Instead of relying on heuristic rules and intense knowledge engineering, these system derive their power from well formulated mathematical formalisms and automatic training procedures. Nevertheless, it is noteworthy that researchers have generally found that performance of these HMM-based systems can be improved when speech knowledge is incorporated, even if only crudely. For example, the use of triphone models conditioned on the left and right neighbors for a given phoneme implicitly models coarticulation, resulting in approximately 50% reduction in word error rate [27].

While it is hard to speculate on what future speech recognition systems would be like, I believe there are many ways current systems can be made more powerful by the proper utilization of speech knowledge. In this section, I will provide two examples in the area of signal representation and feature extraction.

Signal Representation

Current speech recognition systems perform significantly worse than humans on the same task, even under ideal circumstances [10]. When the operating conditions deteriorate, the difference between human and machine performance becomes even more dramatic. There is clearly much to be learned from studying the process by which human listeners decode the speech signal. While little is known about the decoding process beyond the eighth cranial nerve, advances in auditory physiology and psychophysics [15,22,43] have begun to shed some light on the nature of representations of the speech signal in the human peripheral auditory system. As a result of this pioneering work, many researchers have begun to propose speech signal representations that take into account these known properties of the auditory system [31,23,12, 18,44].

In the recognition system under development in our group, the speech signal is first transformed into a representation based on Seneff's auditory model [44]. The model has three stages. The first stage is a bank of linear filters, equally spaced on a critical-band scale. This is followed by a nonlinear stage that models the transduction process of the hair cells and the nerve synapses. The output of the second stage bifurcates, one branch corresponding to the mean firing rate of an auditory nerve fiber, and the other measuring the synchrony of the signal to the fiber's characteristic frequency.

We believe that outputs from various stages of this model are appropriate for different operations in our system. The nonlinearities of the second stage produce sharper onsets and offsets than are achieved through simple linear filtering. In addition, irrelevant acoustic information is often masked or suppressed. These properties make such a representation well-suited for the detection of acoustic landmarks. The synchrony response, on the other hand, provides enhanced spectral peaks. Since these peaks often correspond to formant frequencies in vowel and sonorant consonant regions, we surmise that the synchrony representation may be particularly useful for performing fine phonetic distinctions.

There has been some evidence suggesting that a representation based on auditory modelling can offer performance advantage, especially when the signal is degraded by noise [16,20,8]. Recently, we conducted a set of formal evaluations that compares several different signal representations [33]. To limit the scope of our investigation, we selected the task of classifying up to 16 vowels in American English, using a multi-layer perceptron (MLP) classifier with a single hidden layer [29]. Vowel tokens were extracted from the TIMIT corpus [26]. Training and test sets consist of more than 20,000 tokens (from 500 speakers) and about 2,000 tokens (from 50 speakers), respectively. Three different types of spectral representations were compared, one based on Seneff's auditory model, one based on mel-frequency cepstral coefficients [35], which are very popular among the HMM-based systems, and one based on a cepstrally-smoothed discrete Fourier transform. To strive towards a fair and meaningful comparison, the mel-frequency filters were carefully designed to resemble the critical-band filters of the auditory model. In addition, the dimensionality of the feature vectors was constrained to be equal. Specifically, a 40-dimensional vector, covering a frequency range of 6 kHz, was computed once every 5 msec. The test tokens were either presented to the classifier unchanged, or were corrupted by additive white noise at an averaged signal-to-noise ratio of approximately 10 dB.

Classification performance is summarized in Table 1. For clean testing tokens, the auditory based representations hold a small but consistent advantage over the other representations. When the test tokens are corrupted by noise, this advantage becomes more substantial. These results suggest that the outputs of the auditory model are more immune to noise degradation, and thus will provide better and more robust performance for phonetic classification.

Signal	Classification Accuracy (%)	
Representation	Clean Speech	Noisy Speech
SAM	66.1	54.0
MFCC	61.6	45.0
DFT	61.2	36.6

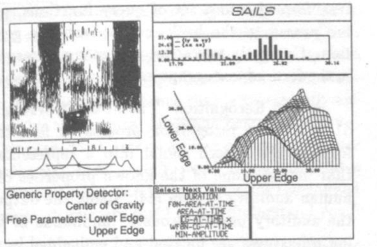
Table 1: Comparisons of vowel classification accuracy (in %) for Seneff's auditory model (SAM), mel-frequency cepstral coefficients (MFCC), and cepstrally smoothed discrete Fourier transform (DFT).

Feature Extraction

Most of the current speech recognition systems do not attempt to extract acoustic attributes that are known to signify phonetic contrasts, but instead use the spectral vectors directly for phoneme and word classification. This choice is partly due to the fact that it is difficult to implement reliable algorithms to automatically extract the acoustic attributes, even if we know qualitatively what they are. For example, there does not yet exist a formant tracker that can determine formant frequencies reliably, especially in regions where the direction and the extent of formant transitions provide important information about the place of articulation for consonants. These algorithms also tend to perform poorly near retroflexed and/or nasalized vowels, making incorrect formant assignment that will lead to catastrophic classification errors.

We have recently experimented with a novel procedure for the extraction of acoustic attributes for phonetic classification. We approach this problem by first defining a set of general property detectors based on our knowledge of acoustic phonetics. We then determine the optimal settings of the parameters by a search procedure, using a large body of training data [39,49]. This procedure is illustrated in Figure 2. In this example, we explore the use of the spectral center of gravity as a general property detector for distinguishing front from back vowels. It has two free parameters, the lower and upper frequency edges. An example of this measurement for a vowel token is superimposed on the spectral slice below the spectrogram, with the horizontal line indicating the frequency range. To determine the optimal settings for the free parameters, we first compute the classification performance on a large set of training data for all combinations of the parameter settings. The results are displayed in the middle-right panel in this figure as a performance landscape, where higher values correspond to better performance. We then search for the maximum on the surface defined by the classification performance. The parameter settings that correspond to the maximum are chosen to be the optimal settings. For this example, the classification performance of this attribute, using the automatically selected parameter settings, is shown at the top right corner. Note that an attribute can also be used in conjunction with other attributes, or to derive other attributes.

We believe that the procedure described above is an example of successful knowledge engineering in which a speech scientist provides the knowledge and intuition, and the machine provides the computational power. Frequently, the settings result in a parameter that agrees with our phonetic intuitions. In this example, the optimal settings for this property detector result in an attribute that closely follows the second formant, which is known to be important for the front/back distinction. Our experience with this procedure suggests that it is able to *discover* important acoustic parameters that signify phonetic contrasts, without



Generic Property Detector: Center of Gravity
Free Parameters: Lower Edge
Upper Edge

Figure 2: An example of interactive discovery of acoustic attributes for phonetic classification.

resorting to the use of heuristic rules.

Do these attributes offer performance advantage over the direct use of spectral information? We recently performed an experiment on a task of classifying 38 phoneme labels using 55,000 and 9,000 training and testing tokens, respectively, from 350 speakers [30]. The input to an ANN classifier is either the spectral vectors from the auditory model plus segment duration (a 241-dimentional vector), or a set of 80 automatically determined acoustic attributes. The performance for the spectral and attribute representations were 72% and 74%, respectively. This result suggest that the use of acoustic attributes can improve classification performance by a small amount, but at potentially considerable computational savings, since the input vector has been reduced by a factor of three.

FROM RECOGNITION TO UNDERSTANDING

Speech Understanding: The Issues

Speech communication among humans is an active process that utilizes many different sources of knowledge, some of them deeply embedded in the linguistic competence of the talker and the listener. For example, utterances such as "let us pray" and "lettuce spray" can presumably be disambiguated based on acoustic-phonetic knowledge alone, which can be determined from the signal. However, distinguishing others, such as "meet her at the end of the street" and "meter at the end of the street" will require syntactic knowledge. Still others, such as "it is not easy to recognize speech" and "it is not easy to wreck a nice beach," cannot be disambiguated without knowledge of discourse context. On the one hand, higher level linguistic knowledge can serve to constrain the permissible word sequences. Thus, for example, the phoneme sequence /**wɛrɪzɪt**/ is more likely to be "where is it" than "wear is it," simply because the first one makes more sense. On the other hand, such knowledge helps us understand the meaning of an utterance, which is essential in spoken language communication. The dual role of *filtering* and *understanding* played by syntactic, semantic, and discourse knowledge enables us to converse freely, and to solve problems jointly using spoken language.

All of the systems reviewed earlier have as their goal the production of an orthographic transcription of what was actually spoken. As long as the proper word sequence is produced by the system, it matters little what the underlying linguistic message is. As a result, linguistic knowledge is utilized only to constrain the search space. The constraints are typically implemented as a statistical grammar that specifies the probability of a word given its predecessors. While these simple language models have been effective in reducing search space and improving performance, they do not begin to address the issue of speech understanding. Indeed, many applications suitable for human/machine interaction using spoken language require a system possessing the capability of solving a problem interactively with a user. In addition to converting the speech signal to text, the computer must also understand the user's request, so as to generate an appropriate response.

Speech understanding systems offer a new set of challenges to researchers, and raise several important research issues. Perhaps the most important one is the integration of speech recognition and natural language processing technology to achieve speech understanding. Researchers in each discipline need to investigate how to exchange and utilize information so as to maximize system performance. In some cases, one may have to make fundamental changes in the way systems are designed. For example, most natural language systems are developed with text input in mind; it is assumed that the entire word string is known with certainty. This assumption is clearly false for speech input, whereby many words are competing for the same time span, and some words may be more reliable than others because of varying signal robustness. Therefore, one many not be able to use existing natural language systems without making some modifications.

Another issue related to spoken language system development is that the system must operate in a realistic application domain, where domain-specific information can be utilized to translate spoken input into appropriate actions. For example, the verb "serve" conveys two entirely different meanings, depending on whether one is discussing a restaurant or a tennis match. Realistic application is critical to collecting data on how people would like to use machines to access information and solve problems. The use of a constrained task also makes possible rigorous evaluations of system performance. Finally, the system must begin to deal with interactive speech, where the computer is an active conversational participant, and where people produce spontaneous speech, including false starts, hesitations, and the use of words and linguistic constructs unknown to the system [46].

An Example: The MIT VOYAGER System

Over the past two years, we have begun a research program on speech understanding system development. In this section, we will briefly describe one such system [50,52].

Task Description For our first attempt at exploring issues related to a fully-interactive spoken-language system, we selected a task in which the system knows about the physical environment of a specific geographical area, and can provide assistance on how to get from one location to another within this area. The system, which we call VOYAGER, can also provide information concerning certain objects located inside this area. The current version of VOYAGER focuses on the geographic area of the city of Cambridge, Massachusetts between MIT and Harvard University, and has a vocabulary of only 350 words. Nevertheless, it can carry on a reasonable conversation with the user and answer a variety of queries, as illustrated in Figure 3.

User:	where can i get some chinese food
VOYAGER:	I know of two Chinese places. (VOYAGER shows them on the map.)
User:	which one is closest to me
VOYAGER:	Where are you?
User:	mit
VOYAGER:	The nearest Chinese place that I know of is Royal East.
User:	how far is it
VOYAGER:	Royal East is about one half miles from MIT.
User:	how do i get there
VOYAGER:	I will try to find the best way to walk there. If you are on the same side of the street as MIT, follow the traffic .....

Figure 3: Example of a conversation between a user and VOYAGER that partly illustrates the system's ability to handle discourse information. All words in the user's queries are represented in lower case, with no punctuation, to underscore the fact that such information is not available to the system.

VOYAGER is made up of three components. The first component, SUMMIT, converts the speech signal into a set of word hypotheses. The natural language component, TINA, then provides a linguistic interpretation of the set of words. The parse generated by the natural language component is then transformed into a set of query functions, which are passed to the back-end for response generation. Currently VOYAGER can generate responses in the form of text, graphics, and synthetic speech.

Speech Recognition The SUMMIT system [49, 51] starts the recognition process by first transforming the speech signal into a representation that models some of the known properties of the human auditory system [44]. Using the output of the auditory model, acoustic landmarks of varying robustness are located and embedded in a hierarchical structure called a dendrogram [17]. The acoustic segments represented in the dendrogram are then mapped to phoneme hypotheses, using a set of automatically determined acoustic parameters in conjunction with conventional pattern recognition algorithms. The result is a phoneme network, in which each arc is characterized by a vector of probabilities for all the possible candidates.

Words in the lexicon are represented as pronunciation networks, which are generated automatically by a set of phonological rules. Probabilities derived from training data are assigned to each arc, using a corrective training procedure, to reflect the likelihood of a particular pronunciation. Presently, lexical decoding is accomplished by using the A^* algorithm [4] to find the best path that matches the acoustic-phonetic network with the lexical network.

Natural Language Our spoken language interfaces make use of a natural language component called TINA [45], which is specifically designed to accommodate the integration of speech recognition with natural language processing. TINA is designed so that its grammar rules and associated probabilities can be automatically trained from a set of correctly parsed sentences. This approach has many advantages, including ease of development, portability, and, most important for use with a speech recognition system, low perplexity. We have, in fact, shown experimentally that grammar probabilities can substantially reduce the perplexity of the resulting language model [45].

The grammar is entered by the developer as a set of simple context-free rewrite rules, which are augmented with parameters to enforce syntactic and semantic constraints. The rule set is transformed automatically to a network form. The parser uses a best-first search strategy. Control includes both top-down and bottom-up cycles, and key parameters are passed among nodes to deal with long-distance movement and agreement constraints. The probabilities provide a natural mechanism for exploring more common grammatical constructions first. TINA also includes a new strategy for dealing with movement, which can handle efficiently nested and chained gaps, and rejects crossed gaps.

Control Strategy The integration of the speech recognition and natural language component is currently achieved using an N-best algorithm [7,47, 52], in which the recognizer can propose its best N complete sentence hypotheses one by one, stopping with the first sentence that is successfully analyzed by the natural language component TINA. In this case, TINA acts as a filter on *whole sentence* hypotheses. If all top N word string candidates fail to parse, then the system provides the canned response, "I'm sorry but I didn't understand you."

Application Back End Once an utterance has been processed by the language understanding system, it is passed to an interface component which constructs a command function in order to generate the appropriate response. Figure 4 gives an example of how a query is transformed into a command function. Note that the functions can be nested to construct more complicated functions. The back-end also has some rudimentary but nevertheless effective discourse capability, so that it can deal with simple anaphora, as well as ambiguous queries, as illustrated in Figure 3.

Query:	Where is the nearest bank to MIT?

Function:	`(LOCATE (NEAREST (BANK nil)`
	`(SCHOOL "MIT")))`

Figure 4: Example of the translation of a query into a command function for accessing the necessary information from the database.

Performance Evaluation In order to evaluate VOYAGER's performance, we collected a corpus of some 5,000 spontaneously spoken sentences from 100 speakers [46]. The system was trained on approximately 70% of the data and tested on 10%. Errors in the system can occur in several ways; the recognizer can mis-recognize a word, the natural language system can fail to generate a parse, an unknown word can appear, or a query can be outside of VOYAGER's domain. All in all, the system could correctly execute approximately 52% of the queries from unknown users [52]. Only 12% of the queries resulted in incorrect response from the system, which can be viewed as catastrophic error. The remaining 36% of the queries prompted the "I'm sorry but I didn't understand you" message. Currently, VOYAGER is implemented on a SUN-4 workstation, using four commercially available signal processing boards, and runs in 3-5 times real-time.

CONCLUDING REMARKS

Why have advances in speech science in general and phonetic science in particular contributed so little to speech recognition research? I believe that one of the primary reasons must be the fact that our knowledge in this area is very spotty. In many areas, we know quite a bit more than we did a few decades ago. However, every shred of knowledge we possess is more than offset by the vast amount that still eludes us. Locally, the jigsaw puzzle is beginning to fit together, but the overall picture is far from clear. For example, we know that phoneme durations can be very important in signifying phonetic contrast. Despite great gains in our knowledge about segment duration, however, we still do not have an adequate durational model that can simultaneously account for variables such as local phonetic context, higher level linguistic constraints, and speaking rate [36]. Without the complete picture, linguistic use of segment duration for speech recognition is likely to meet with only limited success.

I don't mean to sound pessimistic. Quite the contrary, I think there are ways speech knowledge can help to improve recognition performance, and in this paper I have only given a few examples. Over the past decade, there appears to be a gradual polarization in the positions taken by researchers on speech recognition. Some researchers, mostly engineers enchanted by the elegance of mathematics and the power of computing, believe that the problem will be solved if only we can have enough training data. In their view, speech science has very little to contribute to the solution of the problem. Others, mostly speech scientists who had devoted decades to trying to understand human speech communication, scorn the use of statistical modelling. For them, the solution will not emerge until we truly discover the key that unlocks the speech code. Neither of these extremes can possibly be right. While decades may pass before we can develop systems capable of understanding unconstrained spoken language, we are fast approaching a time when real systems with restricted capabilities will begin to emerge. These systems will in all likelihood operate only in limited domains, but will nevertheless help us interact with computers with greater ease and efficiency, thereby making them more accessible to more people. Success in developing these systems will most likely belong to those who can incorporate the acquired knowledge, however incomplete, into a proper computational model, whose parameters can be determined using a large body of data and the vast amount of computing power that is at our disposal.

ACKNOWLEDGEMENT

Much of the work described in this paper was done in collaboration with Jim Glass, Hong Leung, Mike Phillips, and Stephanie Seneff over the past five years. Their contribution is gratefully acknowledged.

REFERENCES

[1] Averbuch, A., L. Bahl, R. Bakis, P. Brown, G. Daggett, S. Das, K. Davies, S. De Gennaro, P.

de Souza, E. Epstein, D. Fraleigh, F. Jelinek, B. Lewis, R. Mercer, J. Moorhead, A. Nadas, D. Nahamoo, M. Picheny, G. Shichman, P. Spinelli, D. Van Compernolle, and H. Wilkens, "Experiments with the Tangora 20,000 Word Speech Recognizer," *Proc. ICASSP-87*, pp. 701-704, Dallas, TX, April 1987.

[2] Bahl, B., A. Cole, F. Jelinek, R. Mercer, A. Nadas, D. Nahamoo, and M. Picheny, " Recognition of Isolated-Word Sentences From a 5000 Word Vocabulary Office Correspondence Task," *Proc. ICASSP-83*, pp. 1065-1067, Boston, MA, April 1983.

[3] Baker, J. M.,"The Dragon System - An Overview," *Proc. Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-23, No. 1, pp. 24-29, February 1975.

[4] Barr, A., E. Feigenbaum, and P. Cohen, *The Handbook of Artificial Intelligence*, 3 vols., William Kaufman Publishers, Los Altos, CA, 1981.

[5] Carlson, R., B. Granström, and L. Nord, "The KTH Speech Database," *Speech Communication*, vol. 9, no. 4, pp. 375-380, August 1990.

[6] Carré, R., R Descout, M. Eskenazi, and M. Rossi, "The French Language Database: Defining, Planning, and Recording a Large Database," *Proc. ICASSP-84*, pp. 42.10.1-4, San Diego, CA, 1984.

[7] Chow, Y, and R. Schwartz, "The N-Best Algorithm: An Efficient Procedure for Finding Top N Sentence Hypotheses", *Proc. DARPA Speech and Natural Language Workshop*, pp. 199-202, October 1989.

[8] Cohen, J., "Application of an Auditory Model to Speech Recognition," *Proceedings, Montreal Symposium on Speech Recognition*, p. 8, July 1986.

[9] Cole, R., R. Stern, M. Phillips, S. Brill, A. Pilant and P. Specker, "Feature-Based Speaker-Independent Recognition of Isolated English Letters," *Proc. ICASSP-83*, pp. 731-733, Boston, MA, April 1983.

[10] Cole, R., R. Stern, and M. Lasry, "Performing Fine Phonetic Distinctions: Templates Versus Features," *Variability and Invariance in Speech Processes*, J. S. Perkell and D. H. Klatt, Eds. Hillsdale, NJ: Lawrence Erlbaum Assoc., 1985.

[11] Cole, R., "Spoken Letter Recognition," *Proc. DARPA Speech and Natural Language Workshop*, pp. 385-390, Hidden Valley, PA, June 1990.

[12] Delgutte, B., "Speech Coding in the Auditory Nerve: II. Processing Schemes for Vowel-like Sounds," *J. Acoust. Soc. Amer.*, vol. 75, no. 3, pp. 879-886, March 1984.

[13] Doddington, G. D., and T. B. Schalk, "Speech Recognition: Turning Theory to Practice," *IEEE Spectrum*, vol. 18, no. 9, pp. 26-32, September 1981.

[14] Dutoit, D., "Evaluation of Speaker-Independent Isolated-Word Recognition System over Telephone Network," *Proc. European Conference on Speech Technology*, pp. 241-244, Edinburgh, Scotland, September 1987.

[15] Fletcher, H., "Auditory Patterns," *Review of Modern Physics*, vol. 12, pp. 47-65, 1940.

[16] Glass, J., and V. Zue, "Signal Representation for Acoustic Segmentation," *Proc. First Australian Conference on Speech Science and Technology*, pp. 124-129, November 1986.

[17] Glass, J. R., "Finding Acoustic Regularities in Speech: Applications to Phonetic Recognition," Ph.D. thesis, Massachusetts Institute of Technology, May 1988.

[18] Goldhor, R., "Representation of Consonants in the Peripheral Auditory System: A Modeling Study of the Correspondence between Response Properties and Phonetic Features," Ph.D Thesis, Massachusetts Institute of Technology, 1985.

[19] Hedelin, P. and D. Huber, "The CTH Speech Database: An Integrated Multilevel Approach," *Speech Communication*, vol. 9, no. 4, pp. 365-374, August 1990.

[20] Hunt, M., and C. Lefebvre, "Speaker Dependent and Independent Speech Recognition Experiments with an Auditory Model," *Proc. ICASSP-88*, pp. 215-218, April 1988.

[21] Jelinek, F., L. Bahl, and R. Mercer, "Design of a Linguistic Decoder for the Recognition of Continuous Speech," *Proc. IEEE Symposium on Speech Recognition*, pp. 255-266, Pittsburgh, PA, April 1974.

[22] Kiang, N. Y-S., T. Watanabe, E. Thomas, and L. Clark, *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve*, Research Monograph No. 35, The M.I.T. Press, Cambridge, MA, 1965.

[23] Klatt, D., "Speech Processing Strategies Based on Auditory Models," *The Representation of Speech in the Peripheral Auditory System*, R. Carlson and B. Granström, Eds. New York: Elsevier/North-Holland, pp. 181-196, 1982.

[24] Kubala, F. and R. M. Schwartz,"A New Paradigm for Speaker-Independent Training and Speaker Adaptation," *Proc. DARPA Speech and Natural Language Workshop*, pp. 306-310, Hidden Valley, PA, June 1990.

[25] Kuwabata, K., K. Takeda, Y. Sagisaka, S. Katagiri, S. Morikawa, and T. Watanabe, "Construction of a Large-Scale Japanese Speech Database and Its Management System," *Proc. ICASSP-89*, pp. 560-563, Glasgow, Scotland, May 1989.

[26] Lamel, L., R. Kassel, and S. Seneff, "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," *Proc. DARPA Speech Recognition Workshop*, Report No. SAIC-86/1546, pp. 100-109, February 1986.

[27] Lee, K. F., *"Automatic Speech Rcognition: The Development of the SPHINX System,"* Kluwer Academic Publishers, Boston, 1989.

[28] Lennig, M., "An 86,000 Word-Recognizer Based on Phonemic Models," *Proc. DARPA Speech and Natural Language Workshop*, pp. 391-396, Hidden Valley, PA, June 1990.

[29] Leung, H. and V. Zue, "Phonetic Classification Using Multi-Layer Perceptrons," *Proc. ICASSP-90*, pp. 525-528, Albuquerque, NM, May 1990.

[30] Leung. H., J. Glass, M. Phillips and V. Zue, "Detection and Classification of Phonemes Using Context-Independent Error Back-Propagation," *Proc. ICSLP-90*, pp. 1061-1064, Kobe, Japan, November 1990.

[31] Lyon, R., "A Computational Model of Filtering, Detection, and Compression in the Cochlea," *Proc. ICASSP-82*, pp. 1282-1285, Paris, France, 1982.

[32] Mariani, J., "Recent Advances in Speech Processing," *Proc. ICASSP-89*, pp. 429-440, Glasgow, Scotland, May 1989.

[33] Meng, H. and V. Zue, " A Comparative Study of Acoustic Representations of Speech for Vowel Classification using Multi-Layer Perceptrons," *Proc. ICSLP-90*, pp. 1053-1056, Kobe, Japan, November 1990.

[34] Mérialdo, B., "Speech Recognition with Very Large Size Dictionary," *Proc. ICASSP-87* pp. 364-367, Dallas, TX, April 1987.

[35] Mermelstein, P. and S. Davis, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-28, no.4, August 1980.

[36] Nooteboom, S. G., "Some Observations on the Temporal Organization and Rhythm of Speech," *These Proceedings.*

[37] Pallett, D., "Benchmark Tests for DARPA Resource Management Database Performance Evaluations," *Proc. ICASSP-89*, pp.536-539, Glasgow, Scotland, May 1989.

[38] Pallett, D., J. Fiscus, and J. Garafolo, "DARPA Resource Management Benchmark Test Results June 1990," *Proc. DARPA Speech and Natural Language Workshop*, pp. 298-305, Hidden Valley, PA, June 1990.

[39] Phillips, M.," Automatic Discovery of Acoustic Measurements for Phonetic Classification," *J. Acoust. Soc. Am.*, vol. 84. S216, 1988.

[40] Poritz, A.,"Hidden Markov Models: A Guided Tour," *Proc. ICASSP-88*, pp. 7-13, New York, NY, 1988.

[41] Rabiner, L. R.,"An Introduction to Hidden Markov Models," *IEEE ASSP Magazine*, January 1986.

[42] Russel, M.,"Recent Results from the ARM Continuous Speech Recognition Project," *Proc. DARPA Speech and Natural Language Workshop*, pp. 397-402, Hidden Valley, PA, June 1990.

[43] Sachs, M., and E. Young, "Effects of Nonlinearities on Speech Encoding in the Auditory Nerve," *J. Acoust. Soc. Amer.*, vol. 68, no. 3, pp. 858-875, September 1980.

[44] Seneff, S., "A Joint Synchrony/Mean Rate Model of Auditory Speech Processing," *Journal of Phonetics*, vol.16, no.1, pp. 55-76, 1988.

[45] Seneff, S., "TINA: A Probabilistic Syntactic Parser for Speech Understanding Systems," *Proc. ICASSP-89*, pp. 711-714, Glasgow, Scotland, May 1989.

[46] Soclof, M. and V. Zue, " Collection and Analysis of Spontaneous and Read Corpora for Spoken Language System Development," *Proc. ICSLP-90*, pp. 1105-1108, Kobe, Japan, November 1990.

[47] Soong, F., and E. Huang, "A Tree-Trellis Based Fast Search for Finding the N-best Sentence Hypotheses in Continuous Speech Recognition," *Proc. DARPA Speech and Natural Language Workshop*, pp. 199-202, June 1990.

[48] Wilpon, J., C.H. Lee, and L.R. Rabiner, "Improvements in Continuous Digit Recognition Using Higher Order Spectral and Energy Features," *Proc. ICASSP-91*, Toronto, Canada, May 1991.

[49] Zue, V., Glass, J., Phillips, M., and Seneff, S. "Acoustic Segmentation and Phonetic Classification in the SUMMIT System," *Proc. ICASSP-89*, pp. 389-392, Glasgow, Scotland, May 1989.

[50] Zue, V., Glass, J., Goodine, D., Leung, H., Phillips, M., Polifroni, J., and Seneff, S. "The VOYAGER Speech Understanding System: Preliminary Development and Evaluation," *Proc. ICASSP-90*, pp. 73-76, Albuquerque, NM, May 1990.

[51] Zue, V., Glass, J., Goodine, D., Phillips, M., and Seneff, S. "The SUMMIT Speech Recognition System: Phonological Modelling and Lexical Access," *Proc. ICASSP-90*, pp. 49-52, Albuquerque, NM, May 1990.

[52] Zue, V., Glass, J., Goodine, D., Leung, H., Phillips, M., Polifroni, J., and Seneff, S. "Recent Progress on the VOYAGER System," *Proc. DARPA Speech and Natural Language Workshop*, Hidden Valley, PA, June 1990.

[53] Zue, V., S. Seneff, and J. Glass,"Speech Database Development at MIT: TIMIT and Beyond," *Speech Communication*, vol 9, no. 4, pp. 351-356, August 1990.