Deployable Automatic Speech Recognition Systems: Advances and Challenges

Biing-Hwang Juang
Robert J. Perdue, Jr.
David L. Thomson

Advances in automatic speech recognition (ASR) technology promise new products and services that capitalize on the enhanced capability of human-machine communication. This paper:

- Addresses several key issues in the design of a deployable ASR,
- Discusses new dimensions and developments of the technology,
- Presents several examples of successful applications of the ASR technology in AT&T's products and services, and
- Describes the research effort needed to realize many new telecommunications applications in years to come.

Introduction

Two of the building blocks of civilization are speech and tools. Yet, while both speech and tools have a long history, only within the past few years have people been able to talk to their tools—and have their tools respond in kind. We may have crossed a great technological threshold, but it is somewhat anticlimactic. Science fiction has presaged such a development for decades, and since people communicate by speech so effortlessly, most do not think of hearing and understanding as a difficult problem.

And yet, no machine today can *hear* and *understand* as a human does, even in simple tasks. It is the purpose of this paper to discuss recent advances and limits of the technology of automatic speech recognition (ASR) and address the issues of making a machine recognize speech successfully in field applications.

Automatic speech recognition is useful in many ways. It can be used for controlling the actions of a machine and for entering and retrieving data. Examples include allowing users to obtain charge account or checking balance information, to move money from one bank account to another, or to order items from a retail catalog. In these examples, a machine that recognizes the spoken word and transcribes it accurately is invaluable.

To be useful, a machine speech recognizer must be *accurate*, *easy to use*, and *cost effective*. First, it must be highly accurate in recognizing human speech. A speech recognition system that does not provide high performance often adds to the user's frustration and may be counterproductive.

Second, it must be easy to use. The more naturally a system interacts with the user—for example, does not require words spoken in isolation—the better the perceived effectiveness by the user.

Third, it must be inexpensive to make. Low cost is essential to place state-of-the-art technology in applications where real values and benefits can be delivered to the user.

In the past few years, AT&T has gained a great deal of experience using speech recognition in products and services. As an example, the AT&T Voice Recognition Call Processing (VRCP) application for the automation of operator calls (0 + number) handles several million domestic calls per day with high accuracy from homes, airports, businesses, and public phones. The VRCP system prompts callers to state the type of call that they want to place and then processes the call accordingly. As another example, the AT&T's 800 Speech Recognition Service recognizes digits spoken directly to the system

Panel 1. Abbreviations, Acronyms, and Terms

ASR—Automatic speech recognition

Cepstral mean normalization techniques identify and remove linear distortion in an ASR signal.

Cepstrum—Fourier transform of the log-spectrum of a signal. Used in ASR to represent the salient properties of speech.

DTMF—Dual tone multifrequency

Fourier transform — An algorithm to perform frequency analysis on a signal.

GPD—Generalized probabilistic descent algorithm used to find the optimal parameter values to minimize the ASR error rate.

HMM—Hidden Markov model (See Panel 2)

Incremental entropy—A measure of the increase in randomness, as well as the amount of information in a signal source. Modifying the signal results in change in the information amount carried in the signal. It can be incremental or decremental.

MAP—Maximum a posteriori adaptation, to take advantage of the general prior knowledge of the model parameters of a word or word segment, for example, and make extremely efficient use of the new data on the speaker to revise the model for an improved recognition rate.

ML—Maximum likelihood model

Minimum classification error formulation, a formulation of the pattern recognition problem in terms of error rate minimization, which has become the basis of discriminative GPD training.

N-best decoding algorithm was developed to allow for incorporating string syntax information, as well as the potential confirmation dialog between the ASR machine and the user.

OSPS—Operator Service Position System, which provides both operator-assisted and automatic support of such "0 +" services as credit card and reverse charging.

Optimal vector quantizer associates an arbitrary vector observation, such as a spectrum, with one of a finite number of "typical" vectors, based on the principal of nearest neighbor. The chosen "typical" vector is more regular than the given arbitrary vector and the index of the "typical" vector is very useful for transmission.

UCS—Universal Card Service

UV—Utterance verification

VIP—Voice Interactive Phone

VRCP—Voice Recognition Call Processing

to route calls to the appropriate location. AT&T products, such as the Intuity™ Conversant® system, use speech recognition in a variety of applications, including inquiring about account balances and using credit cards.

AT&T's experience has motivated the need for features like *key word spotting*, the ability of a recognizer to pick out the required vocabulary word in the presence of extraneous speech or noise; and *barge-in*, the ability of the caller to interrupt the system prompt. These are examples of technical aspects of a speech recognition system that makes it acceptable to callers.

We shall elaborate on AT&T's experiences in this field and discuss the technical issues involving more ambitious applications of automatic speech recognition. From a research point of view, these issues represent many new technological dimensions that are often beyond the conventional approach of treating speech recognition as a problem of linguistics.

This paper begins by presenting recent technological advances in the section "Technical Aspects of Speech Recognition." The next section "Real World Applications," explains several of AT&T's experiences in deploying ASR and provides insights into how to make a maturing technology useful in real applications. The last section, "Further Technological Challenges," discusses some work that lies ahead as we approach the new telecommunications era of the next century.

Technical Aspects of Speech Recognition

Research in the area of acoustic-phonetics in the past few decades has produced a body of knowledge that is often used in guiding the design of an *automatic speech recognition system*. It is, however, the advent of more powerful computing tools and the disciplines of signal processing and statistical methods that have made automatic speech recognition technology practical. With the marriage of these technologies, ever more capable speech recognition systems can be developed to meet the visionary need of the next century.

Figure 1 is a block diagram depicting the framework often followed in the design of a recognizer. The user's speech signal is analyzed and a *speech feature*, that is, a set of mathematical representations that characterize a word or subset of words, is extracted for comparison with stored references in the pattern-matching database. A decision scheme determines the word, or the class of the unknown input, based on the matching scores.

Feature Extraction. As depicted in Figure 1, analysis methods facilitate the process of feature extraction. Almost all the existing systems employ the framework of *short-time spectral analysis*, which computes from the speech waveform a sequence of power spectra, that is, the

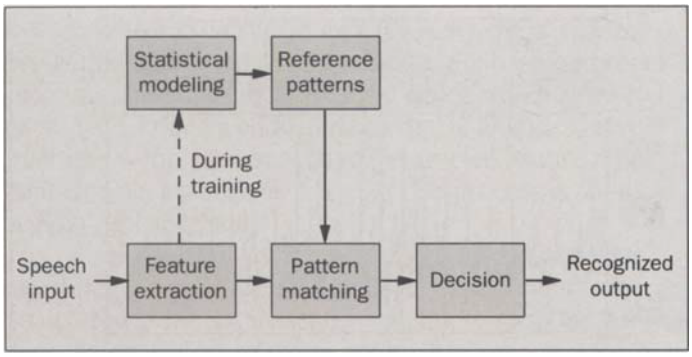

Figure 1. Speech is recognized by analyzing the spoken words and extracting speech features, which are compared against stored references in the pattern matching block. The result is the recognized output of the process. Statistical modeling to produce reference patterns is done during the development phase of the system.

distribution of power or energy at different frequencies. Typically, the speaker's speech is sampled in 10- to 30-millisecond segments, with some overlap between segments.

The short-time power spectrum of each 10- to 30-millisecond speech sample encapsulates the relevant acoustic and phonetic information of the speech signal. The power spectrum undergoes compression, a change of scale, to reflect the auditory perception capability of humans. This eventually leads to the use of *cepstrum*, which is the Fourier transform of the log-spectrum. Such a computational feature extraction process can be considered as a rough approximation to the human auditory system. Experimental results indicate that the cepstrum carries the pertinent linguistic information in a very compact manner and is thus considered a parsimonious representation of each speech segment for the purpose of phonetic identification.

Perceptual studies also indicate that the human auditory system responds to the differences in the spectral sequence more substantially than to the static aspects. This is primarily due to a physiological phenomenon called *short-term auditory adaptation*, which suppresses the redundant sound components appearing repetitively, consecutively in time.

At a very high level, we can illustrate this point by considering the phrase "speech communication." The vowel parts (_ _ee_ _ _o_ _u_i_a_io_) are the relatively static sections of the signal, while the consonant parts (sp_ _ch c_mm_n_c_t_ _n) are carried by sounds that change more dynamically. It is harder to guess what the utterance might be from the vowel parts than from the consonant parts. Another way of considering the phenomenon is to regard the presence of vowels in the spectral sequence as static, or background, that provide the listener with very little information content, while the consonants can be considered foreground information, signaling a perceptible change in a sound segment.

The understanding of short-term auditory adaptation led to the representations of *delta cepstrum* and *delta-delta cepstrum*,[1] that signify the sequential change, or the dynamics, of the cepstral sequence. These representations have been shown to bring about substantial improvements in recognition accuracy when augmented to the original cepstral observations.

Other techniques to improve speech recognition include *cepstral filtering*[2] and *parameter filters*,[3] which attempt to suppress the undesirable variabilities due to measurement noise and artifacts, as well as factors attributable to speaker characteristics, such as the resonance of the speaker's vocal cords, rather than phonetic distinctions.

Statistical Modeling. In the statistical pattern recognition approach, modeling and training refer to the process of choosing the appropriate form of probability distributions for a given word, recognizable sound, or word segment, and then estimating, from a given set of known sample observations, the parameters that define the probability distributions for that word, sound, or word segment.

As an example, consider the word "one." Figure 2a shows a number of signals that are all perceived as the same word, "one." These signals look quite different on the surface. The questions to be answered are:

- 1. What does a typical signal of the word "one" look like?
- 2. How does the signal vary among different speakers without changing the sound identity?
- 3. Given the inevitable variations in individual pronunciation, how can the *likelihood* be determined that a signal actually represents the word?

Figure 2b is an illustration of the typical signal. The variation among speakers can be visualized by comparing the typical signal with the other signals. The technical difficulty is in coming up with a mathematical form that describes the variations.

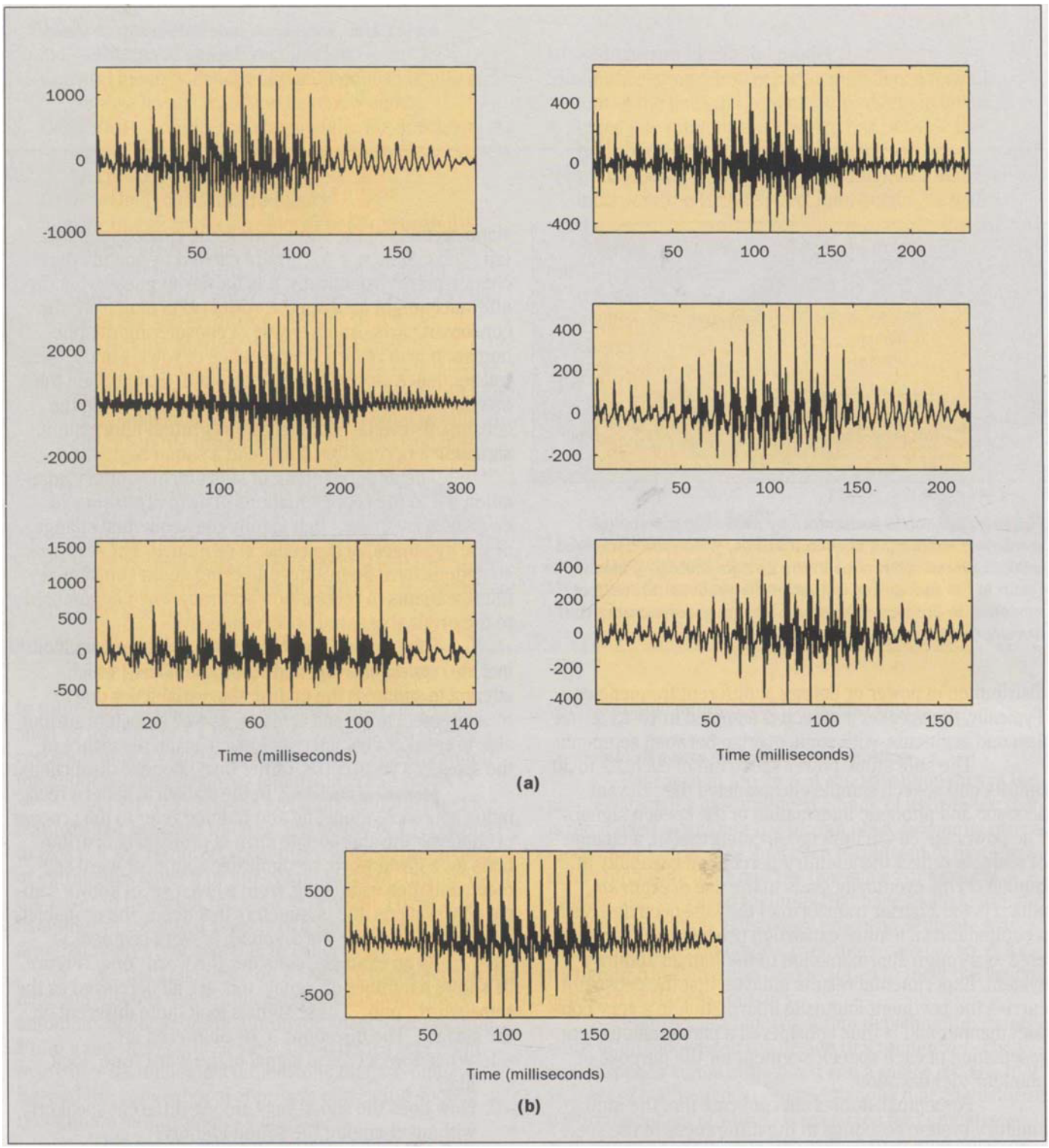

Figure 2. A number of signals are shown in 2a that are all perceived as the same word, "one," although these signals look quite different. The surface differences are due to, among other conditions, the inevitable variations in individual pronunciation. Figure 2b shows the typical signal of the word "one." Using statistical pattern recognition, a speech recognition system can "learn" these variations by choosing the appropriate form of probability distribution for each signal in Figure 2a, and then, from the known sample observation in Figure 2b, estimate the parameters that define the probability distributions for each signal of "one" for future comparisons.

Before statistical modeling was used in speech recognition, traditional approaches to establishing proper databases for recognizing sounds, called *speech references,* relied mostly on the knowledge of linguists, who manually segmented speech and extracted feature representations in a laborious process. This process is often tedious and unreliable.

In contrast, statistical modeling is data driven and, once the prior knowledge in acoustic-phonetics is structured in the database, the training process for the recognition system becomes automatic.

In speech modeling, it is necessary to address two characteristic aspects of the speech utterance, namely, the *local properties* of a phonetic sound, such as the letter "s," "p," "e," "e," and so on in the word "speech," and the *interaction* between adjacent sounds, such as "sp," "pe," and "ee," in the same word. In uttering the word "speech," the speaker's lips are lightly pressed together to create the "p" sound, which affects the perception of the next sound "e" as it is pronounced. By analyzing and identifying the behavior of individual sounds and sound interactions in the same statistical framework, it is possible for a speech recognizer to accurately characterize the probability that the word will be "speech," rather than some other word, word segment, or sound.

The structure of a hidden Markov model (HMM)[4] and its corresponding ability to determine the probability of a word, word segment, or sound are particularly suitable for this purpose. See Panel 2 for a high-level discussion of HMMs.

A hidden Markov model is a probabilistic measure that comprises two levels of probabilistic uncertainties. At the lower level, local characteristics of a sound are described by a distribution of the cepstrum, or speech observations. Such a distribution can be envisaged as a measure governing the generation of the sound in a particular phonetic state. The HMM characterizes how, for example, the "s" sound in the word "speech" is typically pronounced and what variations are expected in the pronunciation of different people.

At the higher level, a Markov chain is used to connect these stochastic states, thereby characterizing the behavior of the sound change. Sometimes, this is called a *doubly stochastic* process.

The common objective of any speech recognition system is to mathematically optimize the model parameters by maximizing the likelihood of the model in producing the given set of known utterances. Given a set of training data for the recognizer, two algorithms, the *Baum-Welch algorithm* and the *segmental k-means algorithm,* provide a mathematical framework to obtain the needed parameter values. These two algorithms usually produce an HMM parameter set that is reasonably good for simple speech recognition applications.

For high performance recognizer designs, two further advances are vital: *discriminative training* and *context-dependent modeling.*

Discriminative Training. Discriminative training[5] was proposed to address the problem caused by the insufficient knowledge of speech distributions. Regardless of the strength of a hidden Markov model, it is an empirical choice and is likely to be different from the "true" speech distribution, rendering the distribution estimation approach suboptimal for the design of a speech recognizer. The discriminative training approach is aimed at directly improving the accuracy of the recognizer, instead of maximizing the likelihood in producing the given utterance. It employs the *generalized probabilistic descent (GPD) algorithm*[5] to find the optimal parameter values. In other words, the objective is not just to describe the typical rendition of a word. Instead, given a particular vocabulary, the object of discriminative training is to determine the best boundary of the utterance of one word versus another word, in order to minimize the probability of an error.

Such methods and their extensions[6] have been shown to substantially reduce the recognition error rate, compared to distribution estimation methods. Many recent products and services have become feasible because of the use of discriminatively trained speech models.[7]

Context-Dependent Modeling. The need of context-dependent acoustic modeling comes from the fact that sounds of the same phonemic identity may possess different acoustic characteristics when spoken in different contexts and, therefore, require separate modeling. An example is "I scream" and "ice cream." Both have the same phonetic identity, but the sound of the "c" in each sample is different enough to indicate different contexts. Without context-dependent modeling, the estimated statistical properties of the phonetic class of a word may not be accurate enough to ensure a high level of recognition.

Context-dependent modeling often involves considerations of structure efficiency—that is, which con-

Panel 2. A Brief Overview of Hidden Markov Models

A hidden Markov models (HMM) permits one to predict the state of a system, that is, whether the system undergoes a change of state, or remains in the same state, over time (t).[16] For example, consider a simple three-state weather system, where once a day ($t+1$) the weather might be in one of three states:

State 1: Rainy
State 2: Cloudy
State 3: Sunny

In this case, we can assume, based on past experience, that the probability of a weather condition changing the next day ($t+2$), depends *only* on the weather condition the day before ($t+1$). Figure 3 shows a diagram of this system. The figure also shows the *transitional probabilities* as labels on the arrows that represent the transitions. For example, according to this model, the probability of having a sunny day tomorrow, given that it is raining today, is 0.3, while the probability that it will continue to rain tomorrow is 0.4. Since the current state of the weather is directly observable (just look out the window), this is an example of an *observable* Markov model.

In speech recognition, however, the state of the speech segment is hidden, that is, not directly observable to the recognition system. (Recall the example of the word "one" shown in Figure 2. The state of the nasal sound sometimes is hard to observe directly and the vast variation indicates that the concept of state needs to encompass the inherent randomness of the signal.) Let's return to the weather system and assume that the observer has no window to the outside world, only the ability to measure the temperature and humidity outside. In this case, the observer can only estimate the sequence of the weather states.

This estimate is based on the knowledge of the transitional probabilities of Figure 3 and the knowledge of the state of the weather as indicated by the temperature and humidity. Although temperature and humidity do not uniquely identify the state of the weather, their statistical distributions depend on it. Therefore, the measurements provide some information about the possible states. Using this information, as well as the transition probabilities and the initial state of the weather, the observer could estimate the entire sequence of weather states ($t+2$, $t+3$, $t+4$, etc.) since entering the room.

In speech recognition, the system uses the HMM to predict the next segment of speech ($t+2$), according to a set of probabilities associated with the speech segment ($t+1$). In reality, such a process would be far more complicated than our simple three-state weather system, since the number of possible states would be much higher than three, and the condition of one state, or speech segment ($t+n$), could depend on the combination of a number of previous states.

texts, or phonemes, warrant explicit modeling. There are 47 phonemes in American English, for example, and that could yield up to 1,209 phoneme pairs (47×47). It is necessary to reduce to a more manageable level the number of models needed by the recognizer to characterize phoneme pairs. Techniques based on such criteria as *incremental entropy* and *frequency of occurrences*[8] have been proposed to appropriate the inclusion of particular contexts for modeling.

For small vocabulary applications, one particularly interesting technique is *hybrid* context-dependency in recognizing connected digits,[9] such as "one," "two," "three," etc. The modeling structure is based on a head-body-tail decomposition of a word, in which the head, the body, and the tail units represent the beginning, the middle, and the ending parts of a word, respectively. Context dependency only exists in the head and the tail units. It was demonstrated that when using this head-body-tail architecture, and with proper discriminative training for the recognizer, an error rate of less than one percent is achievable in recognizing connected digits for many telecommunications applications.[10]

Another technological development is adaptive modeling, which allows for the adaptation of models to a new environment or a new speaker with a minimum amount of training data. One example is adjusting the typical distribution boundaries for recognizing a word or word segment, based on the speaker's unique acoustic-phonetic qualities. Techniques such as *maximum a posteriori (MAP) adaptation*[11] take advantage of the general

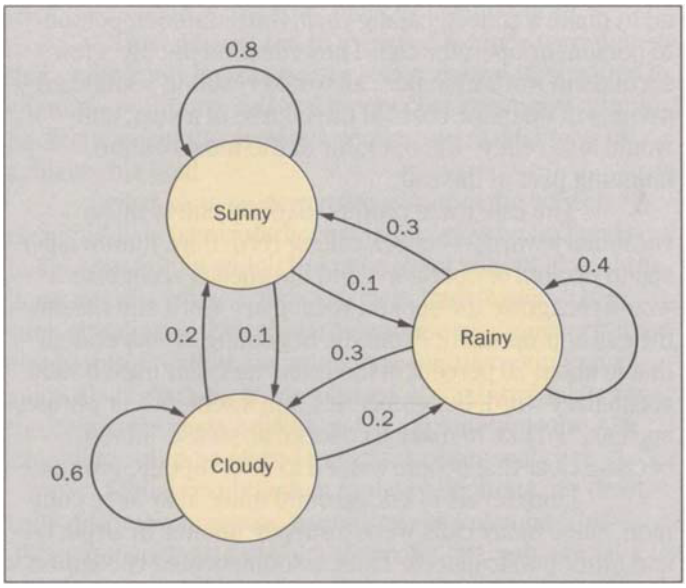

Figure 3. This illustration shows a diagram of the transitions of a simple three-state weather system. The figure also shows the *transitional probabilities* as labels on the arrows that represent the transitions. For example, according to this model, the probability of having a sunny day tomorrow, given that it is raining today, is 0.3, while the probability that it will continue to rain tomorrow is 0.4.

prior knowledge of the model parameters of a word or word segment, for example, and make extremely efficient use of the new data on the speaker to revise the model for an improved recognition rate.

Dealing with Operating Environments. Practical speech recognition often encounters signals with acoustic ambient noise, circuit and equipment noise, and any distortions caused by the transmission channel, including the transducer that converts the speech signal into an electric signal. In addition, speakers may alter their speaking styles under different ambient conditions, such as raising their voices in a noisy environment (that is, the Lombard effects). These undesirable factors can greatly degrade the recognizer's performance. In dealing with these adverse conditions, several methods, based on cepstral processing, have been shown to be effective.[12]

The technique of *cepstral normalization* is a particularly effective method in dealing with this condition. When the adverse condition is mainly due to substantial linear distortion from the channel (such as a mismatch in the microphones used to capture the speaker's response), it will manifest itself in the cepstral domain as a mismatch, or bias, in the cepstral representation. This simplistic distortion model led to the concept of *cepstral mean normalization* or *subtraction*, which identifies and removes the bias.

A *cepstral bias removal algorithm*[13] was proposed to further reduce the impact of the linear distortion. The algorithm is aided by an *optimal vector quantizer* to compute the cepstral bias, which is the difference between the signal being received and a recognized reference point. This estimate of cepstral bias is obtained as a simultaneous result of a better estimate of the undistorted cepstral sequence. This algorithm to remove the general bias (of which the cepstral mean subtraction is a special case) was shown to be able to reduce by 75 percent the performance degradation (in terms of error rate increase) due to mismatch conditions in channels and databases.

Another idea to increase the robustness of a recognizer is to train it under a *mixed-data* (also known as *multi-style*) condition. This is a straightforward approach that provides a huge database to cover as much as possible not only channel conditions, but also a variety of a user's speaking conditions, such as softness, sternness, frustration, and rage. It was shown[14] that, with mixed training data, the GPD minimum error discriminative training method can produce a single set of acoustic models for the recognizer that perform substantially and uniformly better in various test environments than traditional models, known as maximum likelihood (ML) models.

Key Word Spotting and Utterance Verification. One key component in a friendly user interface for speech recognition is to allow the user to speak naturally and spontaneously without imposing a rigid format. In a typical spontaneously spoken utterance, however, there are usually various kinds of disfluency, such as hesitation and extraneous sounds (for example, "um" and "ah"), false starts, and unanticipated ambient noise (for example, tongue clicks and lip smacks).

One approach to accomplishing a natural speech interface, particularly when contemplating domain-specific services, is to focus on a finite set of vocabulary words that are most relevant to the intended task, and then design the system using the technology of key word spotting.

With key word spotting, the user is allowed to

speak freely and spontaneously. The system then detects and identifies the selected in-vocabulary words, or key words, which may be embedded in the user's natural speech, while rejecting any other superfluous words or sounds, including out-of-vocabulary words, ambient sounds, and invalid inputs, such as any form of disfluency or lack of key words.

Utterance verification (UV) is a more general form of key word spotting, in that the detection and identification of key words are accomplished with a certain measure of confidence. Utterance verification can be formulated as a problem of testing statistical hypothesis.

Barge-in. In voice response applications, the system prompts the user to provide speech input. Often experienced users don't want to wait for the prompt to finish before responding. The system should be able to accept the user's input while it is providing the voice prompts and instructions. This technique is generally referred to as barge-in

With barge-in, the recognizer needs to be activated and listening from the beginning of the prompt, while an echo canceler is used to suppress the echoed system prompt. The complication arises from two compounded aspects. One is related to the detection of double talk, when both the speaker and the automated system prompt talk at the same time. The other complication is caused by extraneous sounds, which are likely to be produced either by the user or by something in the background during the long listening period. Either may trigger the recognizer to react.

These problems are solved by properly engineering the echo canceler and providing a robust rejection scheme, which separates reliably the input from the residual signal, since echo cancelers don't always remove 100 percent of the undesired signal.

Real-World Applications

The technological advances elaborated in the last section have made possible many real-world applications of automatic speech recognition. This section highlights these applications, as well as the process of creating the technology in response to real-world problems.

Automatic Call-Type Processing. Starting in 1985 in Reno, Nevada, and continuing throughout the country, AT&T conducted trials to evaluate the feasibility of using speech recognition to determine whether the caller wanted to make a collect, calling-card, third-number, person-to-person, or operator call. The reduction of only a few seconds in work time per call would result in significant savings in operator costs in the course of a year, and would also relieve the operator of the most routine, fatiguing part of the call.

The caller was prompted to say one of those vocabulary words—collect, calling card, third number, person-to-person, or operator—and the speech recognizer would recognize the spoken vocabulary word and handle the call appropriately. From the beginning, it was evident that in about 20 percent of the calls, the caller used a valid vocabulary word, but embedded it in a sentence or phrase, such as, "I'd like to make a collect call, please." It also became clear that people wanted to interrupt the prompt.

High levels of background noise also were common, since many calls were from pay phones in airports and other public places. Thus, a sophisticated recognizer was developed that is capable of key word spotting, permits prompt barge-in, and has a tolerance for high background noise and low-level speakers.

Field trials in Dallas, Seattle, and Jacksonville during 1991 and 1992 led to the deployment of VRCP. Today, the AT&T VRCP platform is integrated in the Operator Service Position System (OSPS) and is deployed at over 30 sites in the AT&T network, and a modified version of VRCP is being used by several Regional Bell Operating Companies. These systems automate more than one billion telephone calls a year with speech recognition.

Interactive Voice–Custom Calling Features. With the recent proliferation of custom calling features, it has become confusing to remember the proper touchtone activation codes, such as the *70 code to cancel call waiting, *66 to activate call forwarding, and the other codes used to control multiple calling features. A trial called Voice Interactive Phone (VIP) was conducted by AT&T and US West in 1992 to help reduce such service barriers. The VIP enhancement requires the customer to remember only one code, which activates a voice response system that lists the options and invites the user to select one by voice.

This capability makes custom calling features much easier to use, as 84 percent of the callers preferred VIP over their present method. It also encourages users to try additional calling features, since 75 percent of the callers tested used the calling features more often—or

tried new services that they had not used before.

The trial pointed to the need for new techniques that make speech recognizers less sensitive to changes in channel conditions. Several newly developed algorithms, as discussed in the previous section, were designed to achieve this goal.

Wireless Voice Digit Dialing. Automatic speech recognition is particularly vital in an "eyes-busy/hands-busy" situation in which the user is not able to look at the input device, or use hands to key in information. One such situation is the cellular telephone in a moving vehicle. In this situation, automatic speech recognition for voice dialing can be accomplished at the terminal, where the cellular phone could have ASR capability, or the ASR capability could be stored in the telephone network.[15]

Signal conditions in mobile telephony are drastically different, however, from those in a standard telephone network. In a cellular environment, substantial interference comes from many sources, such as road and traffic noise, passenger conversation, radios, and multipath fading. The interference may vary in characteristics to a large degree, due to outside traffic and road noise—and whether the vehicle's window is open or closed. Another concern is the articulation of the user, which can vary due to changes in the calling environment—or even the driving conditions.

All these difficulties were overcome by robust recognition methods and by a data collection procedure that accepted spoken digits or key words in lieu of touch-tone input, reflecting the actual operating conditions.

AT&T 800 Speech Recognition. In the AT&T 800 Speech Recognition Service, callers are prompted with a series of choices from a menu and asked to speak a single digit to select an item from the menu. A typical announcement would be "Thank you for calling Economy Airlines. To make a reservation, please say 'one' now; to change an existing reservation, please say 'two' now; to check the status of a scheduled flight, please say 'three' now."

An experienced caller often wants to respond immediately without waiting for the announcement to finish, while novice users may spend up to a half-minute or more listening to menu options before responding. Since such systems are designed to replace the familiar and easy-to-use touchtone signals now used for network access, both novice and experienced users have high expectations of performance, and the system must accommodate them both.

The barge-in capability is critical in such an application. With barge-in, a long response window—the time during which the system will accept a user's response—is provided so that the user does not have to rush to catch the window to respond. Additionally, through a careful analysis of the sounds typically present on the line prior to a valid caller response, a rejection scheme was developed to minimize the incidence of falsely triggering the system.

Another interesting problem encountered during a trial in the United Kingdom was regional dialect. In Northern Ireland, for example, the word "two" is pronounced "tooey." The problem called for the use of sophisticated speech model structures to increase the resolution of the models to recognize these and other speech variables.

Connected Digit Recognition. For many applications of speech recognition over the telephone, accepting connected digit input is essential. Such applications include providing credit card and account numbers and passwords; ordering from a catalog; and, of course, speaking the telephone number.

To ensure high performance, many new techniques discussed in the section "Technical Aspects of Speech Recognition" became imperative. The use of the discriminative training method often cut the error rate by half. Cepstral normalization and bias reduction also recovered a significant portion of the performance that was often degraded by field conditions. And the hybrid (head-body-tail) context-dependent modeling scheme also offered a noticeable boost in the recognizer's accuracy. Furthermore, a flexible *N-best decoding algorithm* was developed to allow for incorporating string syntax information, such as account number syntax or built-in area code constraints that require the middle digit to be either 1 or 0. These new techniques were integrated to provide a recognition accuracy rate greater than 99 percent per word, making many services feasible and acceptable.

One implementation of the connected digit technology is in the Conversant Voice Information System product platform, which has been used in the Universal Card (UCS) 24-hour Customer Services and Information line to handle account information and selective transactions. A Universal Card customer can dial into the 24-hour service number and speak voice commands, in digit form,

as well as the account number, to retrieve information on the card balance and billing address. The service uses the customer's spoken zip code as a verification number to reduce unauthorized access of the information.

Further Technical Challenges

While many recent technological advances have made ASR useful in numerous applications, the challenges ahead are still abundant.

For one, the prospect of having a machine that converses fluently with humans, just as humans do among themselves, still remains a Holy Grail of speech technologists. While pushing ahead to realize this goal, technology advances during the interim must continue to make ASR systems more useful and more beneficial to the user.

Robustness to Ambient and Channel Conditions. The best overall speech recognition system available today is the human auditory system and the human brain. A human's auditory and cortical system has the unique capability of responding to sounds reliably and robustly, via a large redundant neurally connected network. Such a system is much less affected by the ambient channel conditions and by most variations in articulation and pronunciation than the automatic speech recognizers available today. If we could build a machine that mimics the way we ourselves hear, interpret, and understand, we would have a recognizer with accuracy that far surpasses anything developed so far.

It is obvious that the transducers that convert human speech into electrical signals vary tremendously. Telephone handsets vary in types, distortion, spectral shaping, and response level. Microphones are built by a variety of manufacturers and are located at various positions on the handset, with openings of different sizes, and lie in different points within the sound field around the mouth. The use of speakerphones and even ever-changing acoustics of the room further exacerbate the issue of channel variability.

The recognizer must be able to deal with traffic noise, crowd noise, and interference from various other sources. Interestingly, however, much of the noise in a real application originates from the talker. The person who says the word to be recognized also produces a wide range of lip smacks, breathing noises, grunts, throat clearings, and other sounds. These are largely ignored by a human listener, but they can cause problems for a recognizer. Thus, a speech recognizer must be made sufficiently robust to operate and be more economically deployed in a wide range of applications.

Disfluencies in Speech. As we design the recognizer to allow for more natural inputs, many types of disfluencies in speech will substantially affect the performance. Typical disfluencies include hesitation, which often results in superfluous sounds, and repetition of words or partial words. The latter are particularly troublesome, especially in today's recognition systems, in which lexical and language models are incorporated in a rigid manner. Simple rejection schemes or key word spotting methods cannot handle all the complicated disfluencies without the aid of a versatile model for understanding language.

A similar problem, namely out-of-vocabulary speech, also arises in virtually every application of speech recognition. A caller may be asked to say a person's name, but the caller may say instead, "What?" or "Redial" or "713-5283."

It is also common for people to be distracted and start talking to someone else in the room, or even to say nothing at all. The recognizer must be able to accurately determine whether the response is a valid one. As previously noted, this calls for a powerful utterance verification scheme with embedded speech understanding capabilities.

Spontaneous Response. One of the scenarios depicted in the new telecommunications vision is the user conversing freely with the machine in a room environment without holding a handset. Thus, the system presents prompts to the user via loudspeakers. For the user to be able to spontaneously respond or cut in, an *acoustic* echo canceler, instead of a network hybrid echo canceler, is needed to adapt to the room's acoustics, which may change several times during the interchange.

The problem of detecting double-talk will also become more difficult, and the potential distortion introduced by acoustic echo cancellation is expected to be more severe than in the network. Therefore, a better rejection scheme is needed to control the higher potential of triggering a system action due to echo.

Announcement Design. One problem frequently encountered when a new service goes into trial is that the announcement is not as clear to the customers as it was to the designers. Similarly, the call flow may be

found to be unnatural and difficult to follow. This confuses the customer and can cause the caller to speak in such a manner—anger, frustration, or confusion, for example—that ASR performance is actually degraded. Therefore, service trials often are recommended to discover and resolve unanticipated caller behavior.

Conclusion. A broad range of factors must be considered to make speech recognition viable for public services. These factors include user behavior, carefully planned but flexible vocabulary training, service design, robust speech modeling, and the users' high expectations of accuracy and robustness. Through the applications mentioned here and many others, AT&T has made substantial advances in automatic speech recognition technologies and gained a wealth of experience in practical recognition system deployment.

New methods have been introduced that make services using speech recognition respond more reliably to speech input. In contemplating the new communications technology of the next century, the industry faces many more challenges in areas such as robustness, disfluency, and machine dialog/interaction. The work now underway provides AT&T with a solid ground for approaching such a goal.

References

1. J. G. Wilpon, C. H. Lee, and L. R. Rabiner, "Improvements in connected digit recognition using higher order spectral and energy features," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP-91)*, Toronto, Ontario, Canada, May 1991.
2. B. H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass filtering in speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing (ASSP-35)*, Vol. 7, July 1987, pp. 947-954.
3. Bishnu Atal, "From speech to sounds: coping with acoustic variabilities," In Wayne Lea (Editor), *Towards Robustness in Speech Recognition*, Speech Science Publication, Apple Valley, Minn., 1989, pp. 209-220.
4. L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, April 1994.
5. B. H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Transactions–Signal Processing*, Vol. 40, No. 12, December, 1992, pp. 3043-3054.
6. Wu Chou, B. H. Juang, and C. H. Lee, "Segmental (GPD) training of HMM-based speech recognizer," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP-92)*, San Francisco, Calif., March 1992.
7. R. P. Mikkilineni, "Connected digit models using minimum string error rate training for the Conversant System," AT&T Bell Laboratories Internal Memorandum, July 1994.
8. C. H. Lee, L. R. Rabiner, and R. Pieraccini, "Speaker independent continuous speech recognition using continuous density hidden Markov models," In P. Laface and R. De Mori (Editors), *Proceedings of NATO-ASI, Speech Recognition and Understanding: Recent Advances, Trends and Applications*, Cetraro, Italy, 1992. Springer-Verlag, pp. 135-163.
9. J. G. Wilpon and B. H. Juang, "Recent technology developments in connected digit speech recognition," *Proceedings of the International Conference of Spoken Language Processing (ICSLP-94)*, Yokohama, Japan, September 1994.
10. W. Chou, C. H. Lee, and B. H. Juang, "Minimum error rate training of inter-word context dependent acoustic model units in speech recognition," *Proceedings of the International Conference of Spoken Language Processing (ICSLP-94)*, Yokohama, Japan, September 1994.
11. C. H. Lee, C. H. Lin, and B. H. Juang, "A study of speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Transactions–Signal Processing*, Vol. 39, No. 4, April 1991.
12. B. H. Juang, "Speech recognition in adverse environments," *Computer Speech and Language*, Vol. 5, 1991, pp. 275-294.
13. M. Rahim and B. H. Juang, "Signal bias removal for robust speech recognition," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP-94)*, Adelaide, Australia, April 1994.
14. K. Ohkura, M. Sugiyama, and D. Rainton, "Noise-robust HMMs based on minimum error classification," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP-93)*, Minneapolis, Minn., April 1993, pp. II—75-78.
15. D. L. Thomson, J. G. Wilpon, R. A. Sukkar, and D. P. Prezas, "Automatic speech recognition in the Spanish telephone network," *Proceedings of Eurospeech'91*, Vol. 2, September 1991, pp. 957-960.
16. L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of IEEE*, Vol. 77, No. 2, February 1989, pp. 257-286.

(Manuscript approved January 1995)

Biing-Hwang Juang *is a supervisor and distinguished member of technical staff in the Speech Research Department at AT&T Bell Laboratories in Murray Hill, New Jersey. His research interests include speech recognition, coding, enhancement, and stochastic processes. He joined the company in 1982. He has a Ph.D. in electrical and computer engineering from the University of California-Santa Barbara.*

Robert J. Perdue, Jr., *is a technical manager of core technologies in the Voice Transmission Process Department in Global Business Communications Systems (GBCS) in Columbus, Ohio. His group is responsible for developing advanced speech technology features for voice processing products. He joined the company in 1973. He has B.S.E.E. and M.S.E.E. degrees, both from the Massachusetts Institute of Technology in Cambridge.*

David L. Thomson *is technical manager of the Speech Processing Group in the Services and Speech Technology Department of AT&T Bell Laboratories in Napierville, Illinois. His group develops speech recognition and other signal processing technology for Network Systems products. He joined the company in 1984. He has an M.S.E.E. degree from Brigham Young University in Provo, Utah.*