# Speech and Language Processing for Next-Millennium Communications Services

RICHARD V. COX, FELLOW, IEEE, CANDACE A. KAMM, SENIOR MEMBER, IEEE, LAWRENCE R. RABINER, FELLOW, IEEE, JUERGEN SCHROETER, SENIOR MEMBER, IEEE, AND JAY G. WILPON, FELLOW, IEEE

*Invited Paper*

*In the future, the world of telecommunications will be vastly different than it is today. The driving force will be the seamless integration of real-time communications (e.g., voice, video, music, etc.) and data into a single network, with ubiquitous access to that network anywhere, anytime, and by a wide range of devices. The only currently available ubiquitous access device to the network is the telephone, and the only ubiquitous user access technology mode is spoken voice commands and natural language dialogues with machines. In the future, new access devices and modes will augment speech in this role, but are unlikely to supplant the telephone and access by speech anytime soon. Speech technologies have progressed to the point where they are now viable for a broad range of communications services, including compression of speech for use over wired and wireless networks; speech synthesis, recognition, and understanding for dialogue access to information, people, and messaging; and speaker verification for secure access to information and services. This paper provides brief overviews of these technologies, discusses some of the unique properties of wireless, plain old telephone service, and Internet protocol networks that make voice communication and control problematic, and describes the types of voice services available in the past and today, and those that we foresee becoming available over the next several years.*

***Keywords*—*Dialogue management, speaker recognition, speech coding, speech processing, speech recognition, speech synthesis, spoken language understanding.***

## I. INTRODUCTION

The world of communication in the twentieth century was characterized by two major trends, namely person-to-person voice communication over the traditional telephone network and data communications over the evolving data networks, especially the Internet. In the new millennium, the world of telecommunications will be vastly different. The driving force will be the seamless integration of real-time communications (e.g., voice, video, music, etc.) and data into a single network, with ubiquitous access to that network anywhere, anytime, and by a wide range of devices. From a human perspective, the new network will increase the range of communication services to include expanded people-to-people communications (i.e., audio and video conferencing, distance learning, telecommuting, etc.) and people-to-machine interactions (i.e., messaging, search, help, commerce, entertainment services, etc.). These new services will meet the basic human needs for communication, entertainment, security, sense of community and belonging, and learning, and will increase productivity in numerous ways.

In order to understand the role of speech and language processing in the communications environment of the twenty-first century, we first have to look at how things will change as we build out the new network. There are five areas where there will be major changes to the communication paradigm as we know it today.

1) The *network* will evolve from a circuit-switched connection-oriented network with a 64-kb/s connection dedicated to every voice and dialed-up data call to a packet-switched connectionless network based on Internet protocol (IP).
2) *Access* to the network will evolve from narrow-band voice and data to broad-band multimedia integrating voice, image, video, text, handwriting, and all types of data in a seamless access infrastructure.
3) *Devices* connected to the network will evolve from standard telephones and PCs (personal computers) to a range of universal communication devices including wireless adjuncts, mobile adjuncts, appliances, cars, etc. The common characteristic of such devices is that they have IP addresses and can be networked together to communicate over the IP network.
4) *Services* on the network will evolve from simple dial-up voice and data services to a range of uni-

Manuscript received September 27, 1999; revised May 18, 2000.

The authors are with AT&T Labs-Research, Florham Park, NJ 07932 USA.

Publisher Item Identifier S 0018-9219(00)08101-9.

versal communication services including communication, messaging, find, help, sell, entertain, control, storage, and community services. These services will be synergistic with each other and with features of the network that enable them to seamlessly interoperate with all devices and methods of access to the network.

5) *Operations* will evolve from people-oriented processes (which are extremely expensive and highly inefficient) to machine-oriented processes, including natural language voice interactions with computerized agents, self-provisioning of services, web-based billing and accounting, web-based customer care, and automated testing and maintenance procedures.

The new network provides a wide range of opportunities for speech and language processing to become a major component of the telecommunications environment of the new millennium. First of all, the need for speech and audio coding and compression remains high, even as bandwidth increases dramatically to the home, to the office, and in wireless environments. This need remains because the new network offers the opportunity for high-speed streaming of voice, CD-quality audio, and HDTV-quality video, and each of these technologies imposes tight constraints on network performance to maintain high quality with low delay. Coding and compression enable networks to provide high levels of quality at low delays without requiring excessive amounts of network resources.

The second major opportunity for speech and language processing occurs at the access and device levels. Although there is a range of new devices and access methods that will dramatically change the user experience in interacting with the network and its associated services, there exists only a single device that could become the ubiquitous IP access device, and that is the standard telephone. In order for the telephone to achieve this status, several things need to occur so that any IP-based service can work with the ordinary telephone (we call this capability being voice-enabled). First, we need a way to express the voice-enabled service dialogue in a mode compatible with a range of access devices (e.g., a scripting language like VXML). Next, we need a platform for accessing the voice-enabled service via telephone. One such platform, created at AT&T Labs, is called PhoneWeb, for phone access to web services [1]. Next, we need to provide services with user interfaces that are multimodal, namely capable of identifying the service access device and modifying the user interface to work properly with whatever capabilities exist within the chosen access device. Finally, we need to have a high-quality voice interface in order for the service to work properly when accessed using an ordinary telephone. Hence, we need a capability of generating spoken commands from text [text-to-speech (TTS) synthesis], as well as the ability to recognize spoken inputs (speech recognition), understand their meaning (speech understanding), and maintain a dialogue (spoken dialogue) with the user so as to conduct transactions and provide service.

A third opportunity for speech processing is in the area of user authentication. Speaker verification technology can be used as a convenient and accurate method of authenticating the claimed identity of a user for access to secure or restricted services. Speaker verification via spoken voice utterances provides a degree of robustness and reliability that simply cannot be achieved by using conventional logons and passwords.

Finally, in the area of services and operations, the opportunities for speech and language processing are almost limitless. First of all, there is the area of voice command and control, whereby a service or an operation would be initiated via voice commands. Clear examples of such services are voice-activated agents, voice access to information such as movie schedules, airline schedules, etc. A second area where voice processing plays a major role is in communications and messaging systems, where voice signaling can be used to add new people to a teleconference or where TTS can be used to convert a text message (e-mail) to a voice message (voice mail) so it can be heard over a voice communication system with no display capabilities. A third area where voice processing plays an important role is in help or customer care. Here, the voice processing system acts as a surrogate for an attendant or an operator who would handle the query and either provide the desired information or direct the call to the appropriate resource. Perhaps the ultimate use of voice and natural language processing in the new network is the replacement of dial tone (the way we conventionally interact with the telecommunications system today) with voice tone, where we initiate all service requests via voice commands. In this manner we can make calls, access messages, get help, find people, be entertained, etc.

The bottom line is that the telecommunications environment of the new millennium is ripe for voice and natural language processing to be a major component for service delivery and ease-of-use considerations. In the remainder of this paper, we review the basic technologies of voice and natural language processing and show how they have evolved to be used in services in the existing telecommunications network and how they will evolve to be used in new services of the next-millennium network.

## II. Speech and Language Processing Technologies

### A. Speech Coding

Speech coding is a fundamental technology that has existed for more than 60 years, beginning in the 1930s with Dudley's original vocoder [2], [79]. At that time, the goal of speech coding was to provide a compression technology that would enable copper wires to handle the continual growth in voice traffic on the AT&T network. Fortunately, the original need for voice coding never materialized due to the invention of alternate broad-band transport capabilities provided initially by microwave radio systems, and ultimately by optical fiber transport systems. Most recently, the need for speech coding has resurfaced due to the rapid growth in wireless systems (where digital speech coding is essential for handling the ever-growing traffic), and in voice over IP (VoIP) systems, where speech is just one (very important) data type transported over the IP network.

**Table 1** Taxonomy of Speech Coder Types

| Type of Coder | Bit Rate (bits/sample) | Comments |
|---|---|---|
| ***Direct Quantization*** | | |
| uniform PCM | 13 for telephone bandwidth, 14 for wideband speech | highest quality—suitable for teleconferencing |
| companded PCM | 8 for telephone bandwidth | close to perceptually noiseless |
| ***Waveform Following*** | | |
| Adaptive Differential PCM (ADPCM) | 2 – 5 for telephone bandwidth | slight granular noise |
| Code Excited Linear Prediction (CELP) and MultiPulse Excitation (MPE) | 0.5 – 2 for telephone bandwidth | fair to good quality |
| Frequency Domain Coders | 2 – 4 for telephone bandwidth | fair to excellent quality |
| ***Parametric Coders*** | | |
| LPC-based Vocoders, MELP | 0.15 – 0.5 for telephone bandwidth | fair to good quality |
| sinusoidal coders | 0.25 – 1 for telephone bandwidth | fair to good quality |
| waveform interpolation | 0.25 – 0.6 for telephone bandwidth | fair to good quality |

The goal of a speech coder is to compress the speech signal (i.e., reduce the bit rate necessary to represent a speech signal) for either storage or transmission without excessively distorting it. Speech coding is distinct from the more general problem of audio coding in that the primary signal of interest is the speech itself. Other signals (e.g., background noises or music) may be present along with the speech and, therefore, will be compressed and coded along with the speech. However, such signals are generally incidental in speech coding and generally will be disregarded in our presentation.

In this section, we briefly describe some of the fundamental issues in speech coding. We begin with a cursory taxonomy of the various types of speech coders, characterized by their bit rates, and the resulting speech quality. Next, we review some of the peripheral issues associated with telecommunications, such as performance over various types of networks and in noisy environments. We conclude with a brief discussion of open issues.

Speech coders compress speech by analyzing and then quantizing features of the speech waveform in ways that attempt to minimize any audible impairment. The simplest and most widely used coders in standard telecommunications are little more than basic waveform quantizers (called *direct quantization* in Table 1). International Telecommunications Union (ITU) Recommendation G.711 defines two (A-law and mu-law) 8-bit log pulse PCM quantizers. For a variety of input levels, these quantizers maintain an approximate 35-dB signal-to-quantization noise ratio. This noise level is almost inaudible for telephone bandwidth (200–3400 Hz) speech. The speech sampling rate is 8 kHz, yielding an overall coded speech bit rate of 64 kb/s. Virtually all existing telecommunications applications begin with speech coded by this standard. Although this coding rate is more than acceptable for telecommunications, it limits the quality of the speech (to what we call telephone-quality speech) and, therefore, affects the performance of not just speech coders, but also speech recognition systems. An alternative to telephone bandwidth speech is wide-band speech, also known as commentary-quality speech. Here, the bandwidth is 50–7000 Hz, the sampling rate is 16 kHz, and the quantizer is usually 14-bit uniform PCM. The resulting coded wide-band speech not only sounds better than telephone bandwidth speech, but is also more intelligible for humans and works well with modern speech recognition systems.

The next class of coders is known as *waveform following coders*. These coders attempt to reproduce a likeness of the original speech waveform. As a smaller number of speech features are utilized, greater degrees of compression can be realized (with increased levels of distortion). Two principal attributes of the speech that must be preserved (and tracked over time) reliably with such coders are the local pitch (or fundamental frequency) and the local formants (resonant frequencies of the vocal tract). One waveform following method, called adaptive differential PCM (ADPCM) [3], uses a backward adaptive infinite impulse response (IIR) filter that implicitly "tracks" the formants over time. The difference signal between the unquantized speech signal and the one predicted by the prediction filter is quantized. Another waveform following method, called either codebook excited linear prediction (CELP) or multipulse excited (MPE) [4], contains both a formant tracking filter (known as a short-term predictor) and a pitch tracking filter (known as a long-term prediction filter or an adaptive codebook). The short-term prediction filter is based on an all-pole model of the local speech spectrum obtained through a method called linear prediction analysis. In the encoder, the two prediction filters are used to remove all predictable "redundancy" from the speech waveform. What remains is a residual signal. If this residual were used as an excitation signal for the two filters, the original waveform could be exactly reproduced. In order to reduce the bit rate, the residual signal is approximated by a few pulses to form an approximate *excitation* signal. If these pulses are selected sequentially, the coder is a multipulse coder, while if they are selected jointly from a codebook of possible excitation sequences, it is a CELP coder. The selection of appropriate excitation pulses is carried out in a *perceptually weighted* domain, rather than just minimizing the mse in the waveform domain so that the quantization noise is less audible to the listener.

ADPCM coders have been created for bit rates ranging from 16 to 40 kb/s, with 32 and 40 kb/s giving good to excellent quality coded speech. Below these rates, the resulting speech quality is fair to poor. CELP and multipulse speech coders have been created with bit rates from 4.75 to 16 kb/s. The speech quality of these coders ranges from fair to excellent, depending on the particular coder. CELP coders are widely used for wireless applications because of the high degree of compression they achieve.

A third type of waveform coders, called *frequency-domain coders* in Table 1, is based on performing a frequency-domain analysis of the speech waveform [5]. The time-domain waveform is used as input to a filterbank. The outputs of each individual filter (in the filterbank) are critically sampled, resulting in the same number of samples per unit time as the original time waveform. The compression advantage is obtained by exploiting properties of human hearing and by limiting the number of bands actually transmitted. The frequencies that are not transmitted must either be inaudible (be-

cause of masking properties of the human hearing system) or must be recreated using noise excitation at the decoder. Frequency-domain coders have been selected as the basis for a new wide-band coding recommendation by the ITU-T [6].

Another class of coders, called *parametric coders* in Table 1, does not attempt to reproduce an approximation to the original waveform. Instead, this class attempts to produce a signal that sounds like the original by using a parametric model of speech, analyzing the speech to estimate these parameters, and then quantizing only the parameters. The uncoded parameters are insufficient to reproduce the exact waveform, but are capable of producing a signal similar in sound to the original. A variety of parametric models exist, as shown in Table 1 [7]–[9]; these methods are capable of producing speech whose quality is judged to be between fair and good.

*1) Speech Coder Attributes:* Speech coders are characterized by four general attributes: bit rate, quality, signal delay, and complexity. The bit rate is a measure of how much the "speech model" has been exploited in the coder; the lower the bit rate, the greater the reliance on the speech production model. Bit rate can be either fixed or variable. If the number of bits provided by the coder over time is always the same, the rate is fixed. If the number of bits is controlled by the activity of the speech or by the network, the rate is variable.

Quality is a measure of degradation of the coded speech signal and can be measured in terms of speech intelligibility and perceived speech naturalness as measured by formal subjective testing. Perceived speech quality is a function not only of the coder's ability to preserve the speech signal accurately, but also of things like background noise and other acoustic factors. Speech quality is also affected by the transmission system, especially when the bitstream is corrupted by errors or lost entirely for periods of time (as occurs in wireless applications during a fade).

Signal delay is a measure of the duration of the speech signal used to estimate coder parameters reliably for both the encoder and the decoder, plus any delay inherent in the transmission channel. ITU-T Recommendation G.192 provides limits on the amount of delay that is acceptable for real-time conversations. If the round-trip delay is held below 300 ms and there is sufficient echo cancellation, the quality is quite acceptable. Above this delay, there is increasing difficulty in communication. Eventually, a push-to-talk protocol is necessary to facilitate a two-way communication.

Finally, complexity is a measure of computation (and memory) required to implement the coder in digital signal processing (DSP) hardware.

The "ideal" speech coder has a low bit rate, high perceived quality, low signal delay, and low complexity. No ideal coder as yet exists with all these attributes. Real coders make tradeoffs among these attributes, e.g., trading off higher quality for increased bit rate, increased delay, or increased complexity.

*2) Speech Coding Issues:* Several signal-processing techniques have evolved to improve the quality of speech coders in the presence of impairments due to the acoustic environment of the speech, the transmission system over which it is sent, and other factors:

1) speech enhancement methods that attempt to suppress or eliminate background noise, thereby rendering the output speech quality more acceptable to listeners;
2) voice activity detectors (VAD) that attempt to determine whether speech is actually present so as to utilize the channel more efficiently when speech is absent and to avoid trying to code background signals as speech;
3) frame erasure concealment methods, which detect the loss of long sections of the coded speech parameters (due to lost packets, fading channels, etc.) and attempt to interpolate the speech parameters so as to provide some degree of continuity during the lost intervals [10], [80].

Each of these signal processing methods has achieved various levels of success in applied speech coding systems.

### *B. TTS Synthesis*

TTS synthesis technology gives machines the ability to convert arbitrary text into audible speech, with the goal of being able to provide textual information to people via voice messages. Key target TTS applications in communications include voice rendering of text-based messages such as e-mail or fax, as well as voice rendering of visual/text information (e.g., web pages). In the more general case, TTS systems provide voice output for all kinds of information stored in databases (e.g., phone numbers, addresses, car navigation information) and information services (e.g., restaurant locations and menus, movie guides, etc.). Ultimately, TTS could also be used for reading books (i.e., talking books) and for voice access to large information stores such as encyclopedias, reference books, law volumes, etc.

TTS systems are characterized by two factors: the intelligibility of the speech that is produced and the naturalness of the overall message that is spoken. For the past 30 years or more, intelligibility has been the driving factor in building TTS systems, since without high intelligibility, TTS systems serve no useful purpose. As a result, most modern TTS systems are highly intelligible, with formal tests showing TTS word intelligibility approaching that of naturally spoken speech. Significantly less success has been achieved in making the synthetic speech sound natural. Studies have shown that, even with high intelligibility, there exists a minimum level of voice quality that is essential (we call this "customer quality") before consumers will agree to both listen to synthetic speech on a regular basis and pay for the services associated with using the synthetic speech. Hence, the objective of most modern research in TTS systems is to preserve the high intelligibility of the synthetic speech, and at the same time to provide synthetic speech that is customer quality or higher.

*1) TTS Systems:* A block diagram of a typical TTS system is shown in Fig. 1. The first block is the message text analysis module that takes ASCII message text and converts

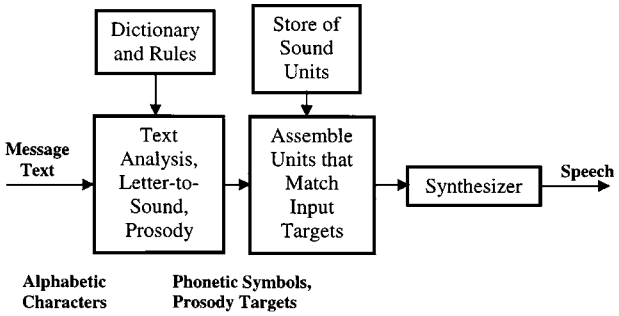


**Fig. 1.** Block diagram of a TTS synthesis system.

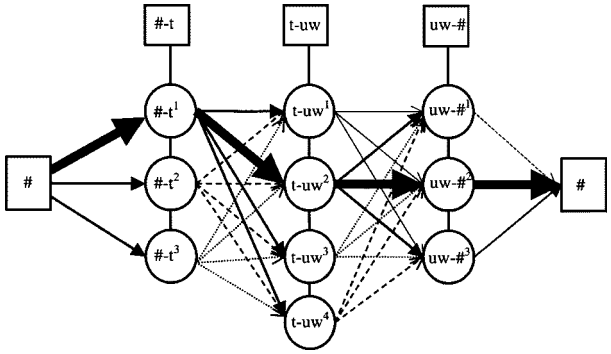


**Fig. 2.** Illustration of unit selection for the word "two."

it to a series of phonetic symbols and prosody (fundamental frequency, duration, and amplitude) targets. The text analysis module actually consists of a series of modules with separate, but in many cases intertwined, functions. Input text is first analyzed and transcribed. For example, in the sentence "Dr. Smith lives on Elm Dr.," the first "Dr." is transcribed as "Doctor," while the second one is transcribed as "Drive." Next, a syntactic parser (recognizing the part of speech for each word in the sentence) disambiguates the sentence constituent pieces in order to generate the correct string of phones, with the help of a pronunciation dictionary. Thus, for the above sentence, the verb "lives" is disambiguated from the (potential) noun "lives" (plural of "life"). If the dictionary look up fails (e.g., for cases like unusual proper nouns), general letter-to-sound rules are used. Finally, with punctuated text, syntactic, and phonological information available, a prosody module predicts sentence phrasing and word accents and, from those, generates targets, e.g., for fundamental frequency, phoneme duration, and amplitude. The second block in Fig. 1 assembles the units according to the list of targets set by the front-end. Then, the selected units are fed into a back-end speech synthesizer that generates the speech waveform for presentation to the listener. For a more general introductory overview of TTS technology, see, e.g., [11].

*2) From Diphone-Based Synthesis to Unit-Selection Synthesis:* During the past decade, concatenative synthesis has been the preferred method in industry for creating high-intelligibility synthetic speech from text. Concatenative synthesis consists of storing, selecting, and smoothly concatenating prerecorded segments of speech after modifying prosodic attributes like phone durations or fundamental frequency. Until recently, the majority of concatenative TTS systems have been diphone-based. A diphone unit encompasses the portion of speech from one quasi-stationary speech sound to the next: for example, from approximately the middle of the /ih/ to approximately the middle of the /n/ in the word "in." For American English, a diphone-based concatenative synthesizer has, at a minimum, about 1000 diphone units in its inventory. Diphone units are usually obtained from recordings of a specific speaker reading either "diphone-rich" sentences or "nonsense" words. In both cases, the speaker is asked to articulate clearly and use a rather monotone voice. Diphone-based concatenative synthesis has the advantage of a small memory footprint (on the order of Mb), since one diphone unit is used for all possible contexts. However, since speech databases recorded for the purpose of providing diphones for synthesis do not sound "lively" and "natural" from the outset, the resulting synthetic speech tends to sound monotonous and unnatural.

Recently, a new paradigm has emerged for obtaining customer-quality TTS. This new method is called unit-selection synthesis. Based on earlier work done at ATR in Japan [12], this new method employs speech databases recorded using a "natural" (lively) speaking style. The database may be focused on narrow-domain applications (such as "travel reservations" or "telephone number synthesis"), or it may be used for general applications like e-mail or news reading. In the latter case, unit-selection synthesis can require on the order of ten hours of recording of spoken general material to achieve customer quality, and several dozen hours for "natural quality."[1] In contrast with earlier concatenative synthesizers, unit-selection synthesis automatically picks the optimal synthesis units (on the fly) from an inventory that can contain thousands of examples of a specific diphone and concatenates them to produce the synthetic speech. This process is outlined in Fig. 2, which shows how the method must dynamically find the best path through the unit-selection network corresponding to the sounds for the word "two." The optimal choice of units depends on factors such as spectral similarity at unit boundaries and on matching prosodic targets set by the front-end. There are two good reasons why the method of unit-selection synthesis is capable of producing customer-quality or even natural-quality speech synthesis. First, on-line selection of speech segments allows for longer units (whole words, potentially even whole sentences) to be used in the synthesis if they are found in the inventory. This is the reason why unit selection appears to be well suited for limited-domain applications such as synthesizing telephone numbers to be embedded within a fixed carrier sentence. Even for open-domain applications, such as e-mail reading, advanced unit selection can reduce the number of unit-to-unit transitions per sentence synthesized and, consequently, increase the segmental quality of the synthetic output. Second, the use of multiple instantiations of a unit in the inventory,

[1]A "natural-quality" TTS system would pass the Turing test of speech synthesis in that a listener would no longer be able, within the intended application of the system, to say with certainty whether the speech heard was recorded or synthesized.

taken from different linguistic and prosodic contexts, reduces the need for prosody modifications that degrade naturalness.

Unit-selection synthesis, as defined in the original CHATR system [12], requires a set of speech units that can be classified into a small number of categories such that sufficient examples of each unit are available to make statistical selection viable. In order to arrive at a robust paradigm (i.e., one that results in consistently high synthesis quality), we have chosen to use half phones as the basic units of synthesis in a way that allows both diphone and phone-based synthesis, and mixtures thereof. This assures a synthesis intelligibility that is comparable to (or better than) that of diphone synthesis with significantly increased naturalness. More details can be found in [13] and [81].

*3) Visual TTS:* In the future, applications such as virtual operators, as well as customer care/help desks on the web, will require realistic "visual agents" that look reasonably human and speak naturally. For these types of applications, lip synchronization of audio TTS and visual TTS (VTTS) is essential. Whether such visual agents are implemented as cartoon-like characters (avatars) using three-dimensional models [14] or synthesized using photo-realistic two-dimensional image technologies (sample-based VTTS) [15], ultimately both approaches will be driven by an MPEG4-standard interface [16].

### C. Automatic Speech Recognition

Innovative speech-controlled user interfaces will unify services in the emerging desktop industry with those available within the traditional telephony industry. For this to succeed, we need to understand and utilize several speech and language technologies for delivering appropriate user interfaces. The most basic of these technologies is *automatic speech recognition* (ASR)—the ability to automatically recognize human speech (on a word-by-word basis). Since humans are able to recognize and understand speech so easily, most people naturally fail to appreciate the difficulties that this task poses for machines. In this section, we review the state-of-the-art in speech recognition while presenting the limitations and challenges that still remain to be solved.

Speech recognition is basically treated as a problem in pattern matching. The goal is to take one pattern, the speech signal, and classify it as a sequence of previously learned patterns, e.g., words or subword units such as phonemes. If speech was invariant to external factors, such as the speaker, the acoustic background, the context, the emotion of the speaker, etc., speech recognition would be a trivial (and solved) problem. However, this is not the case, because the speech signal varies with many factors.

1) *Speaker*—each voice is unique; hence, creating techniques that can accurately and reliably recognize *anyone's* voice and *any* dialect of a given language is a major challenge.
2) *Coarticulation*—the spectral characteristics of a spoken word (or sounds within the word) vary depending on what words (or sounds) surround it.

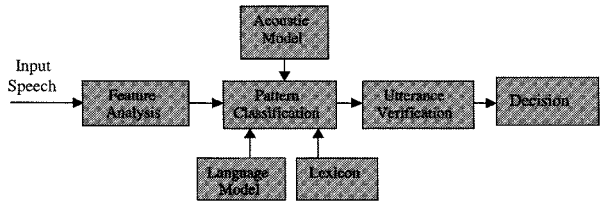


**Fig. 3.** Building blocks of the speech recognition process.

3) *Speaking rate and style*—people speak at different rates and with different pronunciations for the same sounds, thereby making it difficult to get stable patterns for sounds or words that can be used with all speakers and speaking rates and styles.
4) *Environmental conditions*—speech can be difficult to recognize in home environments (background speech from radios or TV), when spoken in a car (road noise distortions), or in noisy backgrounds (airports, train stations).

Each of the above factors contributes some degree of variability to the speech signal. These sources of variability must be carefully considered when developing applications based on speech recognition technology, as it is these characteristics that will ultimately determine whether a speech recognizer will work well in the real world. In addition, people will access the next generation of IP-based services through broad-band digital pipes, using wireless technologies, and in a hands-free mode. Speech scientists are only now beginning to understand and address the difficult problems associated with such diverse methods of access to the network.

*1) Speech Recognition Components:* Fig. 3 shows a block diagram of a speech recognition system. The basic elements of the system consist of the following.

*a) Feature analysis:* The first step in any speech recognition system is extracting, from the input signal, relevant information (i.e., spectral features) that can be used to distinguish one utterance from another. These features must also be computed in such a way as to disregard irrelevant information such as background signals, channel distortion, and speaker dependencies. The greater the ability a speech recognizer has in handling these sources of variability, the more we say that the system is *robust*. Robustness is essential for maintaining a high level of recognition performance across the wide variety of dynamically changing acoustic environments in which a speech recognition system must inevitably operate.

*b) Acoustic modeling:* Currently, most speech recognition systems use statistical models, such as *hidden Markov models* (HMMs), to represent the basic speech patterns (generally referred to as acoustic units) required by the recognizer [17], [18]. Training methods utilize large speech databases and generate statistical models representative of the acoustic units. In the HMM framework, speech is modeled as a two-stage probabilistic process [17], [18]. In the first stage, speech is modeled as a sequence of transitions (the changing sounds) through a directed graph. The states themselves are not directly observable but are represented as observations or features. In the second stage, the features in

a state (the individual sounds) are represented as a mixture of probability density functions over the space of features.

*c) Language modeling and lexicon:* The model of Fig. 3 can, in theory, generate almost an unlimited number of sentences from a large vocabulary of words (the so-called word lexicon). Only a small fraction of these sentences are syntactically correct or semantically meaningful. Since a speech recognition system cannot, *a priori*, determine which sentences are syntactically correct, it is the task of the language model to aid in the discrimination between syntactically likely and unlikely sentences. The most successful and common language model is an $n$-gram word grammar, where the conditional probability of each word in the sentence is a function of the previous $n-1$ words, and the probability of a sentence (a sequence of words) is the product of these conditional probabilities [19], [20]. Large amounts of printed text are generally used as training data to precompute the conditional probability of observed $n$-grams and to estimate probabilities for unseen word subsequences.

*d) Pattern classification:* The heart of any speech recognition system is the pattern classification algorithm, which aligns a sequence of feature vectors from the input utterance to a stored set of previously trained acoustic models. For each input feature vector, the pattern classifier computes the likelihood that the new feature vector was generated from each state in each HMM. The language model probabilities are then combined with the acoustic model likelihoods to compute a score for each possible word sequence. Finally, a decoder searches through the entire space of possible recognition choices to yield the optimum sequence (the one with the highest probability) of words.

*e) Utterance verification and decision:* In order to be able to identify possible recognition errors or nonvocabulary events within the signal, it is important to have a reliable confidence measure of the output of the classifier. If a confidence score is low, this indicates that the best matching sentence is still highly unlikely; in this case the user may be asked to repeat or clarify his input. Utterance verification is generally thought of as being a hypothesis test on the output of the classifier [21]. A measure of confidence is assigned to the recognized string, and the decision box can either accept or reject the hypothesized words by comparing the confidence scores to a decision threshold.

*2) Performance of State-of-the-Art Speech Recognizers:* Over the past several years, many voice-enabled applications have become common in the telecommunications marketplace. At the same time, PC-based software for voice dictation of documents (with unlimited word vocabularies) is being sold by many vendors. Speech recognition technology is now viable in a wide range of applications. In this section, we review current capabilities and performance in the area of unlimited vocabulary recognition.

*a) Unlimited vocabulary speech recognition:* Advances in large-vocabulary speech recognition have been focused in two main areas—creating *search algorithms* that can efficiently scan the space of possible recognition outcomes and developing *acoustic modeling techniques* that more accurately represent input speech and, therefore, improve the raw recognition word accuracy of the system.

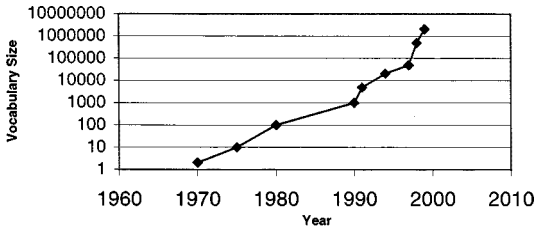


**Fig. 4.** Evolution in speech recognition capability.

Remarkable progress has been made over the past 30 years at developing search techniques that support an ever-increasing word vocabulary size. In 1970, the research community was working hard to develop techniques that could recognize two isolated words (*Yes* and *No*) using minicomputers. Today, we can effectively recognize any size vocabulary in real time on a general-purpose personal computer. Fig. 4 shows the evolution of speech recognition capability over this time period. This growth in capability was made possible by a number of advancements.

1) *Increased speed in computing*—Since 1970, computer speeds have increased by a factor of about 100 000, as they follow Moore's law and double every 18 months. This increase in speed and memory has allowed researchers to experiment with more increasingly complex algorithms.
2) *Improvements in network search algorithms*—Most recognition systems use various types of probabilistic finite-state models to represent the components of the recognition system shown in Fig. 3. For example, $n$-gram word grammars are used to represent language models, multiple-pronunciation dictionaries (derived from the statistics of written language) are used to represent the recognizer's lexicon, and hidden Markov models are used to represent the acoustic models of the speech units. Combining the various finite-state models into a unified network that can rapidly and efficiently be searched has been a major roadblock in building large vocabulary recognition systems until recently. In 1996, Mohri *et al.* presented a framework for efficiently unifying the different finite-state models used in modern speech recognition systems [22], [23]. The result is that techniques now exist for fast (i.e., real time) computation of the optimal path through the network of possible recognition outcomes for virtually any size recognition vocabulary, thus allowing us to break the barrier of vocabulary size and language complexity within modern day recognition systems. This is reflected in Fig. 4, where we see that the slope of the vocabulary growth curve changed radically in 1996.

The current capabilities in speech recognition, in terms of word error rates on some standard corpora, are summarized in Table 2. It can be seen that performance is very good for highly constrained tasks (e.g., digit strings, travel reserva-

**Table 2** Word Error Rates for Several Speech Recognition Tasks

| CORPUS | TYPE | VOCABULARY SIZE | WORD ERROR RATE |
|---|---|---|---|
| Connected Digit Strings—TI Database | Spontaneous | 11 | 0.3% [24] |
| Connected Digit Strings-Mall Recordings | Spontaneous | 11 | 2.0% [25] |
| Connected Digit Strings—HMIHY | Conversational | 11 | 5.0% [24] |
| Resource Management | Read Speech | 1000 | 2.0% |
| Airline Travel Information System | Spontaneous | 2500 | 2.5% [27] |
| North American Business | Read Text | 64,000 | 6.6 [30] |
| Radio recording (Marketplace) | Mixed | 64,000 | 13% [25] |
| Switchboard | Conversational Telephone | 28,000 | 37% [26] |
| Call Home | Conversational Telephone | 28,000 | 40% [26] |

tion) but that the word error rate increases rapidly for unconstrained conversational speech [24]–[27]. It can also be seen that even for a small vocabulary of 11 digits, the digit error rate varies by a factor of 15-to-1 between highly constrained spoken strings (as low as 0.3%) and digit strings embedded in conversational speech, where the digit error rate is 5% [the "How May I Help You" (HMIHY) customer care task described in more detail in the next section]. The *resource management* task was the first large vocabulary recognition task to be uniformly used by all the major speech research labs in the world. For this task, users carefully articulated read speech using a high-quality microphone and a 1000-word vocabulary. Over a four-year period, the error rate on this task was reduced from about 15% per word to about 2% per word. Over the years, National Institute of Science and Technology (NIST) has sponsored progressively harder tasks, with the hardest, *Switchboard*, being a 28 000-word task derived from people having natural conversations over the telephone. At present, the error rate on this difficult task is about 37% per word [25]. Additionally, in 1998, AT&T Labs began experimenting with two very large recognition tasks: a 1-million-word directory information task (i.e., spoken first and last name) and a 460 000 word dictation task whose real-time accuracy is very close to that of the DARPA North American Business News (NAB) task [28], [29].

### *D. Spoken Language Understanding*

The goal of spoken language understanding (SLU) is to extract meaning from the string of recognized words or the set of candidate word strings output by the speech recognizer and to execute an action based on the extracted meaning. Language understanding in spoken dialogue systems typically involves three components: 1) a knowledge representation of the task the system is designed to accomplish; 2) a syntactic analysis of the output of the recognizer; and 3) interpretation of the meaning of the recognizer output in terms of the task representation. This section provides a brief overview of each of these components and then describes three different spoken dialogue systems to provide illustrative examples of different strategies for understanding. For a more comprehensive review of approaches to syntactic analysis, see [31]; for knowledge representation and semantic interpretation in spoken dialogue systems, see [32].

There are many types of knowledge that influence understanding in human–human conversation [33], including:

1) acoustic-phonetic knowledge of the relation between sound and phonemes of speech;
2) phonotactic knowledge of the rules governing legal phonemic sequences and pronunciation variations in a language;
3) syntactic knowledge of the structure of words, phrases, and sentences;
4) semantic knowledge about the meaning of and relationships among words in a sentence;
5) pragmatic knowledge, encompassing knowledge about discourse, the beliefs of the participants in the interaction, the history of the interaction, the task, and general world knowledge.

In spoken-language systems, the first two of these knowledge sources are implicitly embedded in the recognizer's acoustic and language models. As mentioned above, spoken language understanding usually refers to the combination of syntactic analysis, based on grammatical constructs, and semantic interpretation, sometimes also making use of contextual knowledge from the interaction history.

There is a large body of literature on natural language understanding from text (see [34] for an overview). Language understanding systems for text typically generate a parse tree from word strings in order to perform a complete syntactic analysis of the phrase and sentence structure of the input before trying to interpret meaning of the words in a sentence, so that the information afforded by the parse (e.g., part of speech, grammatical dependencies) can be brought to bear for the interpretation. Unfortunately, achieving complete parses of spoken language is often problematic, because of recognition errors and the frequent nongrammatical forms observed in spontaneous speech (including hesitations, restarts, and repairs) [35]. As a result, spoken language understanding systems tend to rely less on complete syntactic analysis than Natural Language Understanding (NLU) systems for text. Fortunately, for many applications, a word-for-word transcription and complete analysis of sentence structure is not required—rather, the task can be completed successfully even if the system only detects keywords and phrases or uses only partial parses of the input [36].

Many spoken dialogue systems restrict their "language" to cover only the limited domain related to the application that the systems addresses, using syntactic information primarily to set constraints on the recognizer output. These constraints can explicitly predefine the set of legal sentences in the application's "language" using a manually constructed grammar, or they can be imposed implicitly, based on the statistical distributions of grammatical forms in labeled corpora that are used to automatically train stochastic grammars [31]. The output of the recognizer, constrained by these grammars, is one or more word string hypotheses. The meaning of the word string is obtained by determining the relationship of the

words in the string to the meaning units in the task's knowledge representation. The following examples illustrate different approaches to spoken language understanding.

*1) Grammar-Based Semantic Specification:* For applications where a handcrafted grammar explicitly defines the set of legal sentences, the syntactic and semantic specifications can be integrated [37]. In these systems, semantic specifications are associated with particular word classes or sentences in the grammar, and when those grammatical constructs are recognized, the semantic specification is returned along with the recognized word or phrase. The semantic specification then results in a call to a function that performs the desired action. For example, in a voice-dialing application, all possible actions (e.g., "call," "fax," "page") may be associated with the semantic specification "METHOD" and the names in the subscriber's voice dialing list associated with the semantic specification "NAME." The recognition result for the utterance "Call John Doe" is returned to the application along with the applicable specification. The text output of the recognizer is parsed into a parse tree containing attribute-value pairs: in the voice-dialing example, the pairs are METHOD: call and NAME: John Doe. Evaluation of the semantic specification NAME generates a database query to retrieve the telephone number associated with the value of NAME (i.e., "John Doe"), and evaluation of the semantic specification METHOD (with value "call") results in initiating a telephone call to the phone number returned in the database query. Thus, the knowledge representation of the task is directly embedded in the recognizer's grammar. Similarly, the structural relationships among the semantic categories are explicitly enumerated in the grammar and associated with desired actions. The interpretation phase consists of initiating the action using the specific values of the relevant semantic attributes. This strategy for implementing "understanding" using handcrafted grammars with embedded semantic specifications is effective for simple applications but results in relatively brittle systems that are unable to handle unanticipated or incomplete input gracefully and that are difficult to port efficiently to new application domains.

*2) Understanding Systems Using Stochastic Grammars:* Spoken language systems that are intended to deal with unconstrained spontaneous speech must be able to process fragmentary input. In these systems, initial syntactic constraints are generally imposed by stochastic language models (described in Section II-C) that are trained on a large corpus of domain-specific utterances. The spoken language understanding components of these systems take the output of the recognizer and focus on understanding only those elements that are critical for task completion, generally ignoring portions of the utterance that cannot be successfully recognized or parsed. In these systems, the constraints imposed by the language model and grammars and the focus on recognizing only task-critical elements combine to yield a more robust system than can be achieved by attempting complete understanding of the input speech. This section describes two understanding systems that work with the output of stochastic grammars.

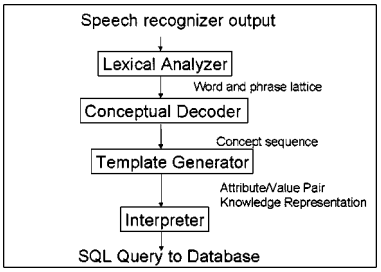


**Fig. 5.** CHRONUS speech understanding system.

*a) CHRONUS:* The CHRONUS understanding system [38], which was first applied to the DARPA Air Travel Information System (ATIS) task, assumes that a spoken utterance is a noisy version of the meaning it was intended to convey, and that an utterance can be modeled by an HMM process whose hidden states correspond to meaning units called concepts. A functional diagram of the CHRONUS understanding system is shown in Fig. 5. The system includes a lexical analyzer that takes the output of the speech recognizer and generates a lattice of hypotheses interpreting the words and phrases according to predefined semantic categories. In the ATIS application, most of the categories correspond to classes of attributes in the air travel database (e.g., cities, numbers, and dates). The lexical analyzer also merges certain word forms (e.g., singular and plural, idiomatic variants of an airport name) to a canonical representation. The conceptual decoder takes the lattice produced by the lexical analyzer and finds the most likely sequence of concepts, by segmenting the sentence into phrases and assigning each phrase to a concept. The set of concepts is predefined, based on the task requirements. Conceptual units for the ATIS task included "destination," "origin," "ground-transportation," and "departure-time," as well as some more abstract units related more to the structure of the sentence than to task-related attributes. Based on the output segmentation produced by the conceptual decoder, the template generator produces a meaning representation of the sentence in a data structure of attribute/value pairs. (Thus the knowledge representation in CHRONUS is the template of attribute/value pairs.) The semantic interpretation phase is rule-based, resolving ambiguities in the template and merging information from previous sentences in the interaction with the current information. The template is then used to construct an SQL query of the travel information database. In CHRONUS, the definitions of the lexical classes, the concepts, and the interpretation rules are handcrafted. The conceptual HMM used in the decoder is trained on a large sample of sentences that were segmented by hand into concepts. In the final formal ATIS evaluation in 1994, the CHRONUS understanding system had the lowest word error rate (6%) [38]. ASR word error rates without the NL component were about 9%.

*b) Task-structure graphs for SLU:* Wright *et al.* [39] described a task-structure graph approach to understanding in spoken dialogue systems. This system has a knowledge representation structure for classification and interpretation that is closely coupled both to the language models used in

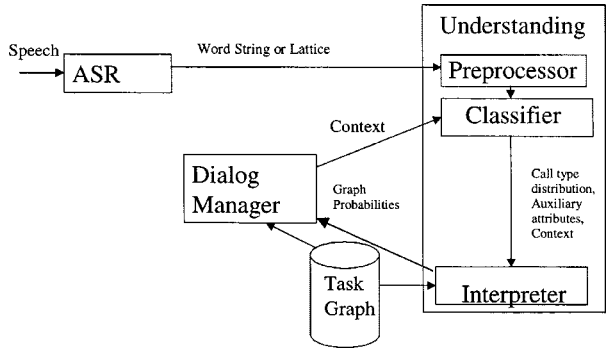


**Fig. 6.** Task-structure graph understanding system.

recognition as well as the dialogue management module. The application addressed in this work was AT&T's HMIHY customer care application. In this call-routing application, the primary role of the SLU component is to categorize the input utterance into one of a small set of categories of call types (e.g., dialing instructions, billing credit, "dial for me," etc.). That call type is then mapped to an appropriate action or response, after the system also extracts any auxiliary information necessary for task completion [40]. In order to accomplish this task, it is not necessary to recognize and understand every word, but only those fragments that are salient to the classification task [41], [42]. Salience is a quantitative measure of the information content of an utterance for a given task. For example, in the HMIHY task, when the phrase "wrong number" is recognized, it is highly likely that the user wants a billing credit for the wrong number and is much less interested in any other possible action, so the phrase "wrong number" is highly salient for call type "Billing Credit."

The method used to accomplish understanding for this task first involves automatically acquiring the salient grammar fragments from a labeled training corpus by modeling the parts of the language that are meaningful to the task, along with their statistical associations to the task. The variable-length salient phrases are automatically generated and selected during the same process that determines the associations between phrases and actions [40]. For robustness and parsimony, these phrases are then automatically clustered into salient grammar fragments, represented as finite-state machines (FSMs). The system then recognizes these phrases in fluent speech by searching the output of a large-vocabulary speech recognizer that uses a statistical language model, which incorporates the FSMs for salient grammar fragments.

The recognized sentence, along with the recognizer's confidence for each recognized fragment and the current dialogue context, are then input to an understanding module, shown in Fig. 6, that consists of a preprocessor, a classifier, and an interpreter [39]. The preprocessor provides a canonical representation of certain classes of words and word strings. The classifier outputs a set of rank-ordered hypotheses for the most likely call type, given the recognized fragments. The knowledge representation used by the interpreter module is a graphical representation of the task that is also used by the dialogue manager in this system. The graph nodes consist of call-type labels, and labels for auxiliary information necessary for completing subtasks, and the graph arcs represent the is-a and has-a relationships between labels [43]. The interpreter performs an inference about the focus of the utterance using the task structure graph, the dialogue context, and the classification of the current sentence. The final result is a set of probabilities for the nodes on the graph, which is returned to the dialogue manager, and which the dialogue manager then uses to determine the appropriate action to initiate. Thus, in this instantiation, the understanding component is coupled tightly with both the structure of the language model used in ASR and with the task structure used by the dialogue manager. This tight coupling makes the system design less modular and, therefore, perhaps less easy to modify for experimenting with new classification or interpretation algorithms. However, in a deployed system, this drawback may be outweighed by the benefits that this coupling among components affords in terms of improved understanding performance.

The use of data-driven techniques to learn the structure and parameters of mapping of salient phrase fragments to classifier categories (for HMIHY) and the mapping of acoustic sequences to concepts (for CHRONUS) has been a significant advance for the field of spoken language understanding, because the resultant systems achieve more robust performance while permitting more natural and unconstrained input. This ultimately results in more successful computer-human interactions. Obviously, these techniques require a large corpus of labeled input to achieve robust estimates of the model parameters and adequate coverage of the language used. For many applications, it is likely that the process for creating an understanding module will be iterative, beginning with a hand-crafted "best guess" at the grammar and evolving the grammar and understanding components as data is collected during system use. This will allow the system to adapt its parameters to better match the language that is actually being used, including tuning the parameters for each dialogue state so that they appropriately reflect the distribution of user responses at any given point in a dialogue. In addition, Riccardi and Gorin [44] have shown that the language of the HMIHY customers changed substantially over time as the system evolved and as users had more experience with the system. This finding suggests that incremental updates of system parameters is a good strategy for continually improving understanding.

### *E. Response Generation*

If spoken language understanding is viewed as the mapping of a string of recognized words into the meaning representation for the task, then response generation can be thought of as the reverse process. That is, the information to be conveyed to the user is held in the data structure containing the knowledge representation of the task, and the response generation component constructs a sentence (or sentences) whose structure is based on the relationships among the task attributes that need to be conveyed, and whose content is the current values of those attributes. For

example, in the ATIS task, a confirmation query about the origin and destination cities might be structured as "Do you want to travel from ORIGIN to DESTINATION?" If the values of ORIGIN and DESTINATION (obtained from the user on previous dialogue turns) are Boston and Chicago, respectively, then the text generated as input to a text-to-speech synthesis system would be "Do you want to travel from Boston to Chicago?" In a system without text-to-speech, response generation may simply be a lookup table of prerecorded prompts for the beginning of the sentence (i.e., "Do you want to travel from"), the origin city, the word "to," and the destination city. For more complex systems where the use of recorded prompts is not an option, customer-quality TTS is required to correctly convey the output of the response generation module. An advanced response generation system could also incorporate knowledge of recognizer certainty, context, and conversational conventions. For example, if the ASR confidence about the term Boston is in doubt, the system could generate the textual sentence with appropriate XML markup for TTS prosody to highlight the system's uncertainty about the origin city. By keeping track of previous output, the response generation system can make appropriate use of pronoun forms and indirect reference, as well as producing natural reductions that typically occur in conversational interactions (e.g., "The first name is Mary. Middle name is Jane. Last name Jones"). For a more detailed review of response generation techniques, see [45].

### *F. Spoken Dialogue Systems*

Spoken dialogue systems extend the functionality of automated telecommunication services beyond simple limited-choice command and control applications to more complex goal-directed tasks that require more than a single dialogue turn between the user and the system. Effective and efficient dialogue systems not only require accurate and robust speech recognition, language modeling, natural language understanding, and speech generation components, but also must incorporate a dialogue management function that oversees the entire interaction.

Fig. 7 shows a block diagram of components of a generic spoken dialogue system. In addition to the speech recognition element discussed above, these systems also include resources for producing and controlling output generation (using recorded prompts or TTS synthesis), accessing databases or other information sources needed to perform the task at hand, and interfacing with the user (in the scope of this paper, telephony control or control of an IP connection). In this figure, there is a separate resource manager to handle low-level coordination of these resources. The role of the dialogue manager is to orchestrate the interaction through the resource manager, including:

1) specifying what information needs to be collected from and delivered to the user;
2) deciding what to say to the user to elicit the required input to the recognizer;

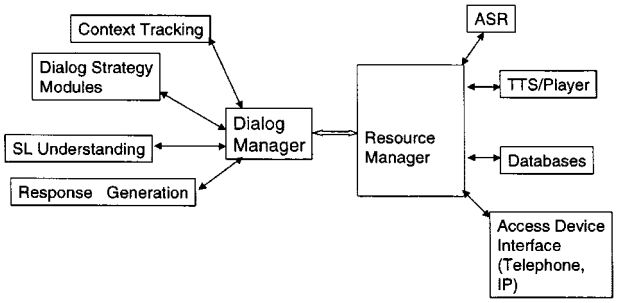


**Fig. 7.** Block diagram of a generic spoken dialogue system.

3) identifying appropriate dialogue strategies to use on subsequent dialogue turns, based on the language understanding module's interpretation of the user's current response and the past context (the dialogue history);
4) determining when to query the database;
5) translating the retrieved information into an appropriate spoken form (response generation).

The roles of the external resources and the language understanding modules have been described in previous sections. This section describes dialogue strategy and the use of context, and also discusses some of the challenges that remain for spoken dialogue systems.

*1) Dialogue Strategy:* To date, the most widely applied metaphor for human–computer spoken dialogue systems has been human–human task-oriented dialogue [46], where it is assumed that the user is interested in accomplishing a task that requires the transfer of various pieces of information from the user to the system and vice versa. Levin *et al.* [47] define a dialogue system as a Markov decision process. In this formalism, the dialogue can be described in terms of a state space, an action space, and a strategy. The dialogue state is all the knowledge the system has of the components it interacts with, including the task representation. Dialogue state changes when an action is taken. The action set is all the possible actions the system can perform (e.g., prompting the user, receiving information from the user, accessing a database). The dialogue strategy specifies the next action to be invoked for each state that is reached during a dialogue session. The dialogue manager is responsible for obtaining the necessary information for accomplishing the task by executing the appropriate sequence of dialogue strategies. Dialogue strategies used in an interaction include subdialogues with the following goals.

1) *Confirmation*—used to ascertain correctness of the recognized utterance or utterances.
2) *Error recovery*—to get the dialogue back on track after a user indicates that the system has misunderstood something.
3) *Reprompting*—when the system expected input but did not receive any.
4) *Completion*—to elicit missing input information from the user.
5) *Constraining*—to reduce the scope of the request so that a reasonable amount of information is retrieved, presented to the user, or otherwise acted upon.

6) *Relaxation*—to increase the scope of the request when no information has been retrieved.
7) *Disambiguation*—to resolve inconsistent input from the user (e.g., "I want to travel at 10 a.m. in the evening").
8) *Greeting/Closing*—to maintain social protocol at the beginning and end of an interaction.

*2) Dialogue Initiative and the Use of Context:* Spoken dialogue systems in use today span a range of flexibility. In the earliest implementations of spoken dialogue systems, the task was often embedded directly in the structure of the application. In these systems, the user was asked a sequence of questions requesting the information elements in a specific order, and the user was obliged to respond with a valid input to each prompt. These systems were often successful for form-filling tasks where only a limited number of elemental attributes were required, only a limited number of values were possible for each attribute, and the dialogue sequence was well known or intuitive to the user. However, these systems were quite brittle to unexpected inputs, often failing in an ungraceful manner. A major advance in spoken dialogue systems occurred when designers of dialogue systems separated the representation of the task (i.e., what information elements are required in order to accomplish the task) from the dialogue structure (i.e., the sequence of dialogue strategies used to elicit the necessary information from the user). This abstraction of the problem into task representation and dialogue strategies not only facilitated reuse of systems for different tasks but also opened up the possibility for mixed initiative dialogue, in which either the system or the user can take control of the dialogue at any point. Because the system always has the representation of all the information that is required for the task, if the user chooses to provide more than one information element in an utterance, the system can make use of everything that the understanding component is able to extract from the utterance. Contrast the following subdialogues from the air travel reservation application.

*System-Initiative Dialogue:*

- *System*: Welcome to the Automated Travel System. Please say the city you want to travel to.
- *User*: Chicago
- *System*: Please say the city you are traveling from.
- *User*: New York
- *System*: On what date do you want to travel from New York to Chicago?
- *User*: May 14th.

*Mixed-Initiative Dialogue:*

- *System*: Welcome to the Automated Travel System. Where would you like to travel?
- *User*: I need to go from New York to Chicago.
- *System*: When do you want to travel from New York to Chicago?
- *User*: May 14th, about 4 in the afternoon.

In the mixed-initiative system, because the user provided both the origin and destination cities, the system skipped ahead to the next missing piece of information (the travel date), streamlining the interaction and making it more natural and conversational. If the user had answered the first question with "Chicago," the system would have asked for the origin city.

Having a separate task representation available and using mixed initiative strategies also allows the system to react to context switches by the user. For example, in the HMIHY customer care task described earlier, the graphical task representation used by both the understanding component and the dialogue manager includes all the information elements and relationships needed for any of the different call-type categories. As a result, if the dialogue manager is proceeding on a path to collect information necessary for making a credit-card call, but the user's next request is interpreted as asking for rate information, the system has the ability to change context to accomplish the user's new goal (i.e., to find out how much the call will cost).

*Context-Switching Dialogue 1:*

- *System*: How may I help you?
- *User*: I'd like to make a credit-card call to area code 973-360-8000.
- *System*: Please say the credit-card number.
- *User*: How much will that call cost?
- *System*: Please hold for rate information.

*3) Challenges for Dialogue Systems:* Mixed-initiative spoken dialogue systems are a reality and are being used for a variety of applications, including information retrieval (e.g., train schedules, traffic, weather, stock quotes), travel planning and ticketing, conference registration, and customer care. However, most of the systems are still limited to a relatively narrow domain. In general, task success rates are lower with mixed initiative systems than with their system initiative counterparts, both because the recognition performance for the less restricted systems tends to be slightly poorer [48], [49] and because novice users who are unfamiliar with the system may be less comfortable with open-ended prompts than with explicit prompting [50]. Obviously, improving robustness of ASR and language understanding will increase the success of mixed initiative dialogue systems. However, Litman and Pan [50] observed that allowing naïve users to switch from mixed-initiative to system-driven dialogue resulted in higher user satisfaction and higher task success, suggesting that automatically detecting when users begin having difficulty and switching to a more system-directed, but "helpful" strategy would result in more successful interactions.

Another factor that could improve the naturalness of interactions would be to maintain and use information from the dialogue history that dialogue theory suggests would be useful. Dialogue theory incorporates the notion that in order to accurately interpret an utterance and respond appropriately, participants in a dialogue recognize the speaker's intentions and beliefs [51]. Plan-based approaches try to infer the beliefs and intentions of the speaker, with the goal of producing actions that correspond to the speaker's plan for achieving the goal. By reasoning about the speaker's intention, these systems have been used for tasks that include handling clarification subdialogues and repairs and generating helpful responses [51]. Currently, theories in the area of dialogue pro-

cessing rely on algorithms for planning and reasoning that depend on well-formed input, and so have been difficult to apply to real-time spoken dialogue systems [52]. As a result, most systems do not represent the user's goals, the system's goals, or many features of the dialogue history, but rather take a strictly structural approach to dialogue, specifying the dialogue as a finite-state automaton (or a stochastic FSA) where each system-user turn constitutes a dialogue state, and transition to the next state is determined by the interpretation of the user's input for the current state. Developing new algorithms for fast and simple identification of a user's goals and incorporating that information in the task representation of today's systems might be a first step toward more natural and intelligent strategies for error detection and recovery or response generation. Similarly, making richer use of dialogue history could also improve interactions. For example, in the HMIHY example above, by maintaining a dialogue history, after the rate goal is satisfied, the system could ask the user whether the previous goal for initiating a credit-card call was still valid, and if so, pick up and complete that dialogue, picking up from the point of the context switch.

*Context-Switching Dialogue 2:*

- *System*: Please hold for rate information. [provides rate]
- *System*: Do you want to make the call to 973-360-8000?
- *User*: Yes.
- *System*: Please tell me the credit card number. . .

Another effective use of dialogue history would reduce the occurrence of sequences of repeated errors. For example, if the user's actions explicitly indicate that the system has misrecognized an utterance, then if that same recognition result is obtained on a subsequent response to the same prompt, the system might anticipate that the same misrecognition has occurred, and modify the dialogue strategy in an attempt to make progress toward the goals, rather than simply repeating the tactic that the user had already flagged as being in error. Use of dialogue history could also improve language generation by allowing the system 1) to match the user's vocabulary in order to mirror the grounding function in human–human conversation [53]; 2) to make appropriate indirect references; and 3) to apply appropriate prosodic contours to indicate which pieces of information in the output utterance are new as opposed to given (i.e., previously provided).

Another challenge that remains for spoken dialogue systems is the lack of a widely accepted evaluation framework. Without such a framework, it is difficult to generalize findings from one system to another. Walker *et al.* [54] have proposed a framework for evaluating dialogue systems that models performance of a dialogue system as a function of what gets accomplished in the interaction (e.g., task success) and how the task gets accomplished (the costs of the interaction). This framework has been applied in several comparative studies of dialogue systems [55].

### *G. Speaker Verification*

Speaker verification is a speech technology that can be used for security and user authentication. Here, the individual claims an identity and the system's task is to accept or reject that claim. Speaker verification provides a reasonably good measure of security for access to a network or a service or to personalized information.

For speaker verification, speech features of the input utterance are analyzed and compared to the features of the claimed speaker. The features most often selected are the same features used for speech recognition. However, for verifying speakers, they are used quite differently. The combination of the speaker's vocal tract and speaking style are what characterize his or her speech, making it distinctive. The vocal tract information and the speaking style are captured in both the instantaneous features and their evolution over time. These features can be compared to statistical models to determine the likelihood of a match from which a decision is made to accept or reject the claimed identity.

Several different types of systems exist for speaker verification. The simplest system conceptually is a fixed password system. After an initial training session (or sessions), the individual makes an identity claim, then speaks his password phrase. The system is conceptually simple because the sequential features should always be similar, as they are spoken in the same order. The features of the utterance are compared to the statistical model for this speaker, and the identity claim is accepted if the features match well and rejected if they are more like those from a background model. The problem with a fixed password system is that it can be thwarted simply by playing a recording of the real individual saying the password phrase. A more secure version of the system can be achieved if the user is asked to speak a given random phase. Ideally, this random phrase, which could be just a string of digits or other words, should include features that are the most characteristic of that speaker.

The remaining question is how well speaker verification technology works in practice. The simplest systems, such as spoken password systems, have the highest accuracy. The performance of any system is measured using a "receiver operating characteristic." There are two types of errors that occur: an identity claim could be falsely verified or it could be falsely rejected. By setting thresholds appropriately, either of these types of errors can be reduced arbitrarily or even eliminated entirely, but at the expense of a large rise in the other type of error. Thus, the entire operating characteristic actually defines the system performance. One particular number that is also of interest is the equal error rate. In practice, equal error rates on the order of 0.5% have been obtained for spoken digit string speaker verification under laboratory conditions [56]–[59]. When a real system is deployed, thresholds must be set *a priori*, although they can be made adaptive during ongoing use of the system. As a result, actual live system performance often has error rates that are two to three times as large.

## III. Networking Issues for Speech and Language Processing

Speech and natural language systems are intended to be essential components of the user interface for services in the

telecommunications network of the twenty-first century. For these interfaces to be successful, they need to work reliably and robustly across a range of access devices and on a variety of networks, including the conventional plain old telephone service (POTS) network, the evolving wireless network, the cable network, and for all types of data (IP) networks. In this section, we review the unique issues that arise when speech and natural language systems are used in the range of environments that will constitute the communications network of the twenty-first century.

### A. Circuit-Switched POTS Telephony

The current circuit-switched telephone network (and the telephone devices connected to the network) introduces limitations that often make speech processing a challenge, including the following.

1) *Reduced signal bandwidth*—A wide-band speech signal contains useful information from about 50 Hz to about 7000 Hz, whereas the POTS network limits the signal bandwidth to 200–3200 Hz, thus removing two important spectral bands of speech information, which often help greatly in recognizing speech.
2) *Speech impairments due to transmission facilities*—The telephone network sometimes introduces impairments that can affect the performance of speech recognition systems. Such impairments include loss of signal level and network echoes.
3) *variable handset quality*—Up until about 20 years ago, all handsets used carbon button microphones. Today, modern day electret microphones are used in almost all handsets. The transmission characteristics for each of these types of microphones are very different, leading to variability in speech recognition.
4) *Accommodation of hands-free environments, especially speakerphones*—Key issues including poor microphone design and acoustic echoes introduced by varying room acoustics and feedback from the speaker to the microphone cause major problems for speech recognition systems and for conversational systems.

Each of these impairments must be carefully handled by the speech processing system for the technologies of Section II to be effective in network applications of speech processing systems.

### B. Wireless Networks

Current estimates for the United States are that there are over 100 million cellular or personal communications services (PCS) phones in use as of mid-2000. It is clear that wireless phones already generate a significant fraction of voice traffic in the network and will continue to do so in the future. Key technologies fueling this growth are speech coding, which enables greater capacity systems, and speech recognition and synthesis, which provide alternative inputs and outputs to compensate for the steadily shrinking keyboards and displays on wireless phones. The marriage of small computing devices, such as personal digital assistants (PDAs), with wireless technology provides a further promising avenue for speech processing technologies.

Wireless networks present some unique challenges to speech processing technologies. The first issue, common to all radio links, is fading. Because of the existence of multiple paths between the transmitter (the cellular phone) and the receiver, the speech signal strength at the receiver goes through periods when it is strong and weak. During strong periods, digital signals can usually be demodulated with no errors whatsoever. During weak periods, bit errors can occur due to interfering signals or general background noise. The bandwidth of the channel, in combination with the ratio of the average signal strength to either the average background noise or the interference level, determines the average capacity of the channel, i.e., the effective bit rate. This impacts the speech coding bit rate that can be transmitted over the channel—hence, the perceived quality of wireless speech. It also implicitly impacts speech recognition performance on the speech transmitted over wireless channels.

The second issue with wireless networks is scarcity of bandwidth. The more users that can share the fixed bandwidth of the network, the less cost there is and the greater the capacity of the network. Therefore, all digital wireless networks use speech coding to increase their capacity. As noted previously, speech coding to low bit rates (order of 4 kb/s) also tends to degrade the performance of speech recognizers.

In summary, the challenges of wireless generally are how to achieve greater capacity in a limited resource (bandwidth) and how to deal with an imperfect channel. For speech processing technologies, the first challenge puts a stress on low-bit-rate speech coding and feature extraction for speech recognition. The second challenge results in a requirement for robust performance on noisy channels for any speech technology. Wireless has been growing at a fast rate for some years, making it a significant fraction of all telecommunications traffic. Thus, it is essential for speech processing technology to perform robustly for wireless channels. Owing to their small size and limited channel capacity, wireless devices and services provide great opportunities to utilize speech processing technology.

### C. Internet Protocol (TCP/IP)

IP is the dominant packet protocol for data in use today. In the future, it is expected to become the dominant protocol for transporting voice and other multimedia data. Transmission control protocol (TCP) is the protocol to ensure that the entire data stream is successfully transmitted over the network (end-to-end). TCP is a "best-effort" protocol, meaning that when data packets are lost, it is assumed they will be retransmitted and eventually will successfully be transmitted and received.

IP voice systems essentially compress speech and chop it up into packets for transport over a TCP/IP data network. Each individual packet encounters a range of switching points (usually routers) on its path from the source to the destination, and different packets from the same speech stream can traverse the IP network over different routes,

depending on the traffic at each of the switching points. This leads to three generic problems with VoIP systems.

1) *Variable transit time through the network*—Each packet can, in theory, take a completely different path through the network. Different paths mean different transit times through the network. Even when every packet takes the exact same path, there is variable transit time since the packet processing time at each switching point depends on local traffic at the time the packet reaches the router.
2) *Generally larger delays than a circuit-switched network*—This is due to the need for compressing the speech and packetizing it into speech packets. Hence, rather than processing the speech on a sample-by-sample basis, as is done in the conventional POTS network, frames of speech are compressed and packetized, thereby leading to extra delays for buffering speech, compressing the buffered speech, packetizing the compressed speech, and the inverse operations at the destination.
3) *Lost packets when there is congestion in the network*—This is due to the nature of TCP/IP, which is a best effort delivery mechanism on top of the IP protocol. There are times when congestion at a router or switch is so high that the buffers cannot handle the overflow traffic. In such cases, entire packets are lost with the intention of having the source retransmit the packet at a later time. For speech, this retransmission mechanism serves no value since a lost packet represents a gap in a "real-time" signal that needs to be filled with some type of signal processing on the received packets.

It will always be necessary to include mechanisms in the speech processing system to deal with each of these data networking problems.

## IV. Communication Services Using Voice and Natural Language Processing

We stated at the beginning of this paper that there were unlimited opportunities for speech and natural language processing to play a major role in the telecommunications environment of the twenty-first century. In the previous sections, we reviewed the capabilities of several speech and natural language processing technologies and listed some of the limitations imposed by means of access to the network. In this section, we discuss the role that speech and natural language processing have already played, the role that these technologies are playing today, and the role we foresee in the future.

There are several market forces that play key roles in the success of future telecommunication-based services, including:

1) the convergence of the computer and telephony industries is moving forward at an ever-increasing pace;
2) people are getting *overloaded* by technology; they want a *simpler and easier* lifestyle.

Given these factors, what will differentiate telecommunications services in the twenty-first century are innovative and compelling user interfaces that create a "media center," unifying emerging desktop services with traditional telephony services. This will allow people to utilize the services they want, whenever they want, and from wherever they are. Such user interfaces will lead to services that are "more intelligent." As networks automatically learn their customers' needs and desires, and customers invest their time to learn how to use these intelligent user interfaces, a strong, "sticky" bond will be formed. Automatic speech and speaker recognition, spoken language understanding, and text-to-speech synthesis technologies will play a critical role in achieving the vision of a simple-to-use intuitive user experience.

We can think of applications based on speech technologies as being divided into several categories.

1) *Command/control*—For these applications, speech technologies are used to replace or augment other types of input modalities, e.g., touch-tones, to control simple applications, for example, extending simple interactive voice response (IVR) applications with the ability to "press or say 1 if you want hardware" or call handling automation via commands like "collect call please." The deployment of speech technologies for these services is mostly driven from the standpoint of reducing the cost of offering the service or making the service more versatile.
2) *Information access*—For this class of applications, speech technologies are used to automatically (without human involvement) access information that would otherwise not be possible if speech technologies did not exist, e.g., getting stock quotes or movie reviews, accessing e-mail, or getting directions over the phone. Another class of information services involves obtaining access to anything that is on the web via voice commands, e.g., "Get me the latest CNN stories on Bill Clinton." These types of services are generally motivated by being able to create new revenue generating opportunities.
3) *Customer care*—As new services or products are offered to the marketplace, the costs associated with providing customer care or help desk functions keep increasing. For this class of applications, speech technologies are being used to replace human agents, thereby automating many of the service support functions and reducing the overall cost of the service.
4) *Audio indexing*—There are a large number of hours of archival speech and video existing in the world today in such forms as movies, TV shows, voice mail, documentaries, and music. In order for the information in these databases to be truly useful, there must be easy-to-use search methods for accessing the information. For example, users should be able to query a database of such material by speaking commands such as "Get me all the news stories on Kosovo from January 1st till today." For textual information, there are several highly effective search engines in existence today. The need for searching audio information using voice or text commands is spawning a new service industry

called *audio indexing*. Information retrieval (IR) technologies are being combined with powerful speech recognition engines to *listen to* and efficiently search audio databases to give customers the information they desire [60]–[63].

In the remainder of this section, we review the voice-enabled services that have been created over the past several years and look at the new services that will exist as a result of new capabilities in speech and natural language processing.

### *A. Voice-Enabled Services—Yesterday*

Services based on speech technology are on their way to becoming a billion-dollar industry. Today, billions of telephone calls each year are being routinely automated based on speech technology. Millions of copies of speech recognition software for dictating letters and memos are sold annually in retail stores by companies such as IBM, Dragon, and Lernout & Hauspie—rivaling that of the best-selling computer games. Speech technologies have truly come a long way.

It is important to bear in mind that speech technologies, despite all the advances, are not perfect. Therefore, most of the early voice-enabled services in telecommunications are those that had the following characteristics:

1) *simplicity*—the resulting services have been easy to use;
2) *evolutionary growth*—the early applications have been extensions of existing systems, such as utilizing voice input to supplement touch-tone data entry for interactive voice-response systems;
3) *tolerance of errors*—given that every speech recognizer, speech synthesizer, or speaker verification system makes occasional errors, the applications must be "fail soft"—i.e., they must be able to gracefully recover from recognition errors.

Several systems that meet the above criteria are described below.

*1) Anser—The First Deployed Voice-Enabled Service:* It has been almost two decades since the first widespread deployment of automatic speech recognition was introduced in the telephone network. In 1981, NTT combined speech recognition and synthesis technologies in a telephone information system called *Anser*—Automatic Answer Network System for Electrical Requests [64]. This system provides telephone-based information services for the Japanese banking industry. Anser is deployed in more than 70 cities across Japan serving over 600 banks. Currently, more than 360 million calls a year are automatically processed through Anser, bringing in about $30 million in revenue to NTT annually.

Using a 16-word lexicon consisting of the ten Japanese digits and six control words, this speaker-independent isolated word speech recognition system enables customers to make voice inquiries and to obtain information through a well-structured dialogue with a computer over a standard telephone. At last report, about 25% of the customers chose to use the ASR capabilities, with a reported word recognition accuracy of 96.5% [65].

Anser provides a win–win scenario for both the service provider and the customer. From the customer's standpoint, the cost of obtaining information about bank accounts is low (approximately the cost of a local telephone call). Also, because most banks are on the Anser network, there is uniformity in accessing banking information across the banking industry. Therefore, customers can access any bank computer using a consistent set of procedures. For the banking industry, Anser allows the banks to provide a much needed service to its customers without having to hire large numbers of people or invest heavily in extra hardware.

*2) Automation of Operator Services:* There are two classes of applications that have spurred the growth of voice-enabled services within the telecommunications industry—those that led to reducing costs of currently offered services and those that created new services and, therefore, new revenue opportunities. Far and away the bigger opportunity has been in the area of cost reduction, mostly notably in the area of automating operator services.

In 1989, Bell Northern Research began deploying Automated Alternate Billing Services (AABS) through local telephone companies in the United States, with Ameritech being the first to offer this service [66]. For this application, ASR technology was used to automate the back-end of "collect" and "bill-to-third-number" calls. After the customer placed a call, a speech-recognition device was used to recognize the called party's response to the question: "You have a collect call. Please say yes to accept the charges or no to refuse the charges."

In 1992, AT&T deployed its first service using speech recognition technology to automate a portion of calls originally handled by operators. The introduction of this service, called *voice recognition call processing* (VRCP), greatly reduced operator workload while increasing the overall efficiency of operator handled calls. The exact task deployed was the automation of the billing functions: *collect, calling card, person-to-person, operator-assisted*, and *bill-to-third number*. Customers were asked to identify verbally the type of call they wished to make without directly speaking to a human operator. A simple five-word vocabulary (the function names) was designed, built, trialed, and deployed from 1986 to 1992. This service is today considered successful not just from a technological perspective but also from a business point of view. From a technology point of view, key advancements in speech recognition technology were achieved to support the needs of the service. For example, both word-spotting and barge-in technologies, commonly used in most new voice-enabled services, were perfected and introduced for the first time in VRCP [67]. From a business perspective, VRCP currently handles more than 1.3 billion recognition attempts per year, more than *all* voice-enabled services in the world put together, with only a 0.3% word error rate. This level of performance has led to savings for AT&T of over $300 million per year or well over $1 billion since its introduction in 1992. The introduction of the VRCP service marked the beginning of the use of voice-enabled services for the mass market.

*3) Credit-Card Account Entry:* There is a whole class of telephone-based services in which users must first identify themselves to the system by entering their account numbers. Usually these services ask customers to enter their account number via touch-tone input. However, there is a large percentage of customers (about 25%) who cannot perform this operation (i.e., they do not have a touch-tone phone) or have problems entering their account number via the touch-tone user interface, and, therefore, must wait for a live agent to come on the line. This increases call holding times and requires more live agents to be available to service the account. In both cases, this means a more expensive service. Several companies have been proactive in trying to reduce costs by using a speech recognition system to allow customers to speak their account number. In 1994, AT&T Universal Card Services customer care help line upgraded their service to use speech technology. Customers now get a prompt asking them to speak or touch-tone their account numbers. Today, millions of calls a years (approximately 60 million calls in 1999) are automated using this feature with a raw string recognition accuracy of over 97% and close to perfect string accuracy after doing appropriate database "dips" to determine whether a recognized string corresponds to an assigned user account in the database.

*4) Reverse Directory Assistance:* In 1993, Ameritech deployed a service called *automated customer name and address* (ACNA) [68]. In this service, customers were provided with name and address information associated with a particular telephone number. After the user provided a telephone number using touch-tone input (currently no speech recognition technology is being used), a search was made in a reverse directory database and text-to-speech synthesis was used to return the desired information to the user. Nynex trialed a similar service in 1992 [69]. For these types of voice-based information access services, where the number of responses that the system must provide the user is extremely large, it is infeasible to record each message, store it, and provide a mechanism to enter new information and change existing information. Therefore, TTS capabilities are an essential component for this service to be successful.

### *B. Voice-Enabled Services—Today*

Yesterday's voice-enabled applications typically provided voice interfaces to existing IVR services or to applications that could have been implemented on an IVR system. Today, however, voice-enabled systems are handling applications where IVR either is not viable or is too cumbersome for one or more of the following reasons.

1) The number of choices available to the user is very large (e.g., tens of thousands of stock names in a stock quote service, or hundreds of cities and airports in an air travel application).
2) The complexity of the interaction would require a deep and bushy IVR menu. Traversing extensive menus using a touch-tone keypad can be quite cumbersome, and, for usability, the number of options at any point in an IVR dialogue must be kept small (typically no more than five) and the nesting of options should not be too deep. High-performance spoken language systems that permit natural speech input allow the user to bypass the rigid hierarchies imposed by DTMF menus.
3) The terms the user is likely to use to describe what he/she wants may not match well to the terms the systems offers in its menu choices. The use of advanced spoken language systems can reduce the memory load on the user by eliminating the need to remember the mapping between what the user wants to accomplish and the specific vocabulary used in the system's prompts or their corresponding DTMF sequences.
4) IVR methods for specifying what the user wants (e.g., by spelling) are ambiguous.

As a result of improved ASR performance and advances in the design of dialogue systems, there are a variety of applications working over the telephone today, including advanced communication services like voice dialing and messaging, as well as services providing information access and transaction management. Speaker verification technology is also incorporated in some transaction and information access systems to ensure secure access to private accounts or information. This section describes a sampling of the voice-enabled applications currently available.

*1) Voice Access to People/Voice Dialing:* One of the biggest opportunities for speech technologies in telecommunications is voice dialing. Currently, to reach an individual, we must have access to the phone number for everyone we want to speak to. Even worse, we have to remember multiple numbers (e.g., home phone, cell phone, business phone, beeper, etc.) for many people and, therefore, need to know when to dial each number. Since telephone numbers are often ten digits long, and with new area codes being introduced at an ever increasing rate, it is becoming impossible to keep track of all the phone numbers of friends and colleagues. Modern speech technologies provide a way to obtain both the convenience of calling people by name (rather than number) and to utilize the power and capabilities of modern telecommunication networks. For broad acceptance, this voice dialing capability must:

1) be *universally* available from everywhere (i.e., from home, office, pay phones, cell phones, PCs);
2) be available for all calls (i.e., POTS, wireless, cable, and IP telephony);
3) work under all conditions;
4) work from a single voice dialing list of names (i.e., a common address book).

Unless the above conditions are met, there will be no incentive for customers to change their dialing habits and use voice dialing for all calls. This is evidenced by previous voice dialing offerings from Nynex, AT&T, Bell Atlantic, Sprint, and others. Such systems have been deployed in only one environment, e.g., only for calls from home, only for calling-card calls, or only for calls from cellular phones. Although these systems have been shown, for the most part, to perform well from the speech technology standpoint (name accuracies greater than

95%), they have had very limited market success because of this lack of ubiquity of access to the voice dialing capabilities. Despite this lack of success with practical systems, voice dialing remains a potential killer application within the telecommunications industry.

*2) Unified Messaging:* Voice-enabled access to unified messaging is another application that we believe will grow into a major business for AT&T and others. People want to have ubiquitous access to their messages, including e-mail, voice mail, and fax (and eventually video-mail). When accessing messages from the desktop, unified messaging is a relatively easy service to provide. However, when users are away from the desktop and want to access the message store via telephone, then speech technologies are essential for providing the most effective and useful service. For example, text-to-speech synthesis is required for reading e-mail to users. Speech recognition is necessary for providing a simple voice-controlled interface to messaging functions, e.g., "Play (or read) my next message" or "Delete this message," and for indexing voice mail for easy retrieval, e.g., "Do I have any messages about my meeting with Larry tomorrow?" Voice control of messaging (and voice dialing) functionality from a customer's cell phone has the additional advantage of alleviating the potentially dangerous situation that currently exists when a user has to take his or her eyes off the road to access the keys on a touch-tone pad in order to enter commands.

In the marketplace, we are beginning to see services using speech technologies that provide remote access to e-mail and/or voice mail. General Magic, partnering with Excite, offers *mytalk*[2] for remote access to e-mail over the telephone using TTS to subscribers of their e-mail system. To cope with the fact that users may have multiple voice and e-mail accounts to contend with, AT&T is currently prototyping services that allow customers to easily voice-enable their *current* voice mail or e-mail services without having to sign up for new messaging services.

*3) Information Access and Transaction Applications:* Automated IVR systems that use DTMF menus to specify queries, retrieve information, and initiate transactions have been deployed by large companies for many years. This section describes a variety of automated information and transaction services that are currently available over the telephone only because of the existence of high-quality voice-enabled interfaces.

*a) Brokerage services:* In the brokerage industry, several large corporations have deployed spoken language systems that allow customers to make trades, review accounts and orders, and obtain market indexes and stock, mutual fund, and options quotes. Currently, these systems span a range of dialogue capabilities. Some systems provide only a system-driven prompting strategy, while other systems allow the user more flexibility and control over the interaction.[3] These systems have been very effective at reducing the percentage of calls that require a human agent, as well as reducing the time it takes users to get information. For example, Fidelity Investment's system is designed to handle an average of 250 000 calls per day, and Schwab's stock quotation system handles more than 50 000 calls per day.

*b) Extended banking services:* As might be predicted, given the success of the ANSER application described in Section IV-A, the banking industry has also started to embrace more advanced spoken dialogue systems. The functionality provided in banking applications includes retrieving account balances, transferring funds, paying bills, and making general inquiries about loan rates and other services. Many of these services are also provided using IVR systems, but by allowing natural speech input, the spoken language systems provide more efficient access to the particular information the customer wants, as well as offering access to customers who do not have DTMF telephones. (DTMF penetration is not as widespread in some countries as it is in the United States)

*c) Inventory and tracking systems:* Systems where long strings (containing both numbers and alphabetic symbols) are used to specify model numbers, customer identifiers, or tracking numbers are also problematic for touch-tone input, because it is difficult to encode these complex strings unambiguously using the touch-tone keypad. Although correct recognition of alphanumeric strings is also a difficult task for speech recognition, performance is very good for strings of known length and composition, verified with checksums or against a database of "valid" numbers. As a result, speech-enabled applications for these tasks are more usable than their touch-tone counterparts. Sample applications in this domain include:

1) the UPS automated package tracking application, where the customer speaks the alphanumeric tracking number to get information about the delivery status of a package;
2) Hewlett-Packard's Y2K Compliance customer information service, where the customer speaks the catalog number of the software product to find out about that product's Y2K compliance;
3) General Electric's Price and Availability Line, which provides customers with current information on price and availability of parts and products.

Such systems are more usable than their touch-tone counterparts.[4]

*d) Travel information and reservations:* The set of information access and transaction applications in the travel domain that are currently deployed or in trial is expanding rapidly, including airline reservations and ticket purchase (e.g., United Airlines' internal system for their employees[5] and Andersen Consulting's internal Via World Network reservation system, train timetable information [70]–[72],

[2]http://mytalk.com.

[3]Audio examples of several different brokerage applications are available on the web; for a demonstration of E*Trade's Telemaster service, see http://www.speechworks.com/customers/customer_listing.cfm.

[4]For audio clips of the Hewlett-Packard systems, see http://speechworks.com/customers/customer_listing.cfm; for descriptions of the UPS and General Electric systems, see http://www.nuance.com.

[5]http://speechworks.com/customers/customer_listing.cfm.

rental car reservation systems [73], and access to local information, including weather, area restaurants information, and sports (the Bell South Voice Access Link).

The American Airlines Dial-a-Flight system exemplifies how adding a spoken language option can enhance the capabilities of an IVR-based application. The dialogue in this system is structured to be completely system driven, asking the user for flight number and the date that the flight information is needed ("today, tomorrow, or in the future"), one piece of information at a time, in a fixed order, while asking the customer to respond after a "beep" tone. This task could be accomplished with an IVR system, but the spoken dialogue system allows the service to be extended to users who may not know the flight number by including a subdialogue that asks for the cities of origin and destination, in order to deduce the flight number from information the user does know.[6]

*e) Directory query :* Directory assistance applications have been of interest in the telecommunications arena for many years, primarily because of the potential cost savings to be realized from automating such services. In the past five years, several field trials of automated telephone directory assistance have been reported. All these systems use system-driven dialogue initiative, with prompts tailored to elicit a single information element per turn. These trials ranged from recognition of localities [74], [75] to the recognition of residential names in a database of about 25 million listings [76]. Performance for locality recognition in three different areas (Quebec, Denver, and Kentucky) ranged from 72% to 80% concept accuracy with 1% false acceptance [75]. The vocabulary sizes for the locality task ranged from about 300 to 800 phrases. For the residential task, 45% of the calls that could potentially be automated in that trial were successfully automated, demonstrating the difficulty of the task. The reasons for failure included recognition errors; user "misbehaviors," including the use of natural language utterances and not knowing some of the information requested by the system (e.g., locality or first name); and adverse acoustic or transmission conditions [76]. It is notable that the speech recognition performance on well-formed utterances in this task was 75% for the field trial, where it had been 92% in the laboratory, highlighting the discrepancy in performance between systems tested under controlled laboratory conditions and real-world usage. This effect has been noted previously in our discussion of robustness in speech recognition systems.

Systems providing voice-enabled directory query for large corporate databases have also become more prevalent over the past five years. Their primary function has been to provide call redirection to the appropriate extension, after the caller dials into a general corporate number.

More complex dialogue systems for directory tasks that allow more natural queries about other attributes associated with a person in a corporate directory (e.g., fax number, e-mail address, location), are currently being prototyped [77]. As mentioned in Section II-C, real-time recognition of very large vocabularies ($>$1 million entries) is currently feasible. However, accurate speech recognition is only one part of the problem in large directory information access. The system must be able to deal with the inherent ambiguities of the directory task. These ambiguities include homonyms (e.g., if the caller says "What's John Lee's fax number?" did he mean John Lee, Jon Li, or John Leigh?), as well as multiple people with the same name (e.g., three different employees, all spelling their name John Lee). In a spoken dialogue system, both these ambiguities must be resolved to provide the caller with the information she or he wants. The homonym ambiguity can be resolved by asking the caller to spell the name. The multiplicative ambiguity must be resolved by asking about other disambiguating information that the system knows and that the system expects the caller will also know (for example, the work location). Effective and quick disambiguation strategies will be necessary for user acceptance of these more complex directory services as they become available in the near future.

*4) Communication Agents:* In the last five years, several services offering voice-activated communication agents or personal assistants have appeared, including Wildfire,[7] Webley,[8] and General Magic's Portico.[9] These services typically combine a set of applications, including voice dialing from a personal address book, outbound messaging, message retrieval (both voice mail and e-mail), call routing, personal calendar, and information retrieval applications (e.g., news, stock quotes, weather, traffic, specialized directories) in a package customized for the user. The agent serves as an intermediary between the subscriber and people (or entities) who are trying to contact him/her. The subscriber is provided with both telephone and web-browser access to the service, making these applications among the first to recognize that customers are likely to access these applications over various access devices at different times, and that some functions, like setting up a personal address book or personal calling priorities and profiles, may be cumbersome to administer without a persistent display. These applications are also cognizant of the need for synchronization of multiple personal information sources, allowing uploads from personal information management systems and access to multiple voice mail and e-mail systems. Currently, voice-activated access to the communications agent is limited to the telephone, but it is easy to envision a future where speech input is also available with the web interface.

### C. Voice-Enabled Services—Tomorrow

As we begin the twenty-first century, and as speech and language processing capabilities improve dramatically, we see a number of key voice-enabled services that will help shape the telecommunications environment. These services include:

1) *agent technology* to manage communications and messaging, to get, index, classify and summarize information from the web (e.g., movie reviews, news), to provide personalized services (to user requirements) and

[6]For an audio clip of this application, see http://www.nuance.com.

[7]http://www.wildfire.com.

[8]http://www.webley.com/index_set.html.

[9]http://www.generalmagic.com/portico/portico_home.shtml.

customized services (to user needs), and to adapt to user preferences;
2) *automated customer care attendants*, which replace IVR systems with interactive service;
3) *call center automation* using natural language voice dialogues for booking airline tickets, car rentals, form filling, catalog ordering, etc.;
4) advanced computer-telephony integration with access to active user registries to aid in completing communications, user caches for frequently called parties, adaptation to use (automatic addition of new names and access numbers), and voice browsers for surfing the web;
5) *voice dictation systems* for generating text responses to e-mail received via a voice interface.

We describe several such systems, as we believe they will be realized, in the remainder of this section.

*1) Customer Care and Call Routing—"How May I Help You":* A major opportunity for speech technologies to change dramatically the way people interact with services is automation of functions such as *call routing* and operations such as *customer care*. *Call routing* services are those in which a customer calls into a central location, then either by speaking with a live agent or interacting with an automated touch-tone menu, has their call routed to a person or automated service that can best handle their needs. *Customer care* operations are those for which customers have a question or problem with a product or service and need help. Billions of dollars per year are spent by corporations to offer such services to their customers via 800-number access services; most of the cost supports live customer agents. In fact, the IVR industry is a multibillion-dollar industry that was, in some sense, created to automate and reduce the costs of offering these types of *call routing and customer care* services. However, touch-tone-based services have many limitations (outlined in the previous section) that a speech-based user experience should be able to overcome. In the next generation of automated customer care services, customers will interact directly with machines, expressing what they want using spoken natural language. By *natural*, we mean that customers will not be restricted as to what they can say, or how they say it; the voice-enabled service will understand what was said and take appropriate action. This type of service represents a paradigm shift over the current generation of IVR interfaces, as well as an extension in scope compared to most speech-enabled applications deployed today.

In 1996, AT&T Labs created a prototype system, called HMIHY with the greeting *"AT&T. How May I Help You?"* [39], [40], [42]. It combines research in speech recognition, spoken language understanding, dialogue management, and text-to-speech synthesis to give users a more natural user interface to the task of automation of operator services. (The dialogue aspects of the HMIHY system were described above in Sections II-D and II-F.) This system was successfully evaluated on over 20 000 live customers in the AT&T Network in 1997. For this service, customers called into an AT&T operator services office and, instead of speaking with a live agent, were greeted with the voice prompt "AT&T. How may I help you?" The automated system then carried on a dialogue with the customer to try to determine how best to either route the call (to another service or an attendant) or how to respond to the user request for service. We have described the dialogue aspects of this HMIHY system earlier in this paper.

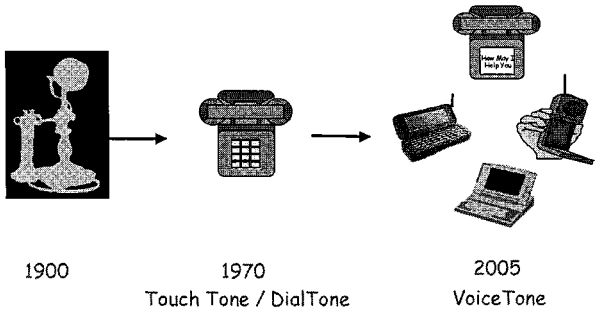


**Fig. 8.** VoiceTone, the future paradigm for communications.

Based on an analysis of many hours of recorded customer voice queries to the system, it was found that customers for this service generally needed some type of help on about 15 different topics. Such a service was found not to be viable using traditional IVR technology, since only five options can be given to the customers at each menu prompt, and, therefore, it generally took multiple prompts before the customer had access to the desired feature or service. In actual service, it was shown that an IVR-based service led to only 84% of the customers getting routed correctly after 40% of the customers had bailed out to an operator. By allowing customers to speak naturally and in conversational speech, as opposed to using touch-tones, about 90% of the customers could be routed correctly with only 10% being falsely transferred to an operator. Clearly, these results show the opportunity speech technologies have for providing a much higher quality of customer care while, at the same time, reducing the costs of providing the service.

*2) VoiceTone:* As shown in Fig. 8, at the beginning of the twentieth century, in order to make a call or get some service, a customer would pick up a telephone and speak to a live telephone agent, or operator, whom we call *Mabel*. Mabel was friendly, helpful, understood what the customer wanted, and could provide a high level of intelligent individualized service; Mabel was the local version of an "intelligent agent." Since there were no "buttons" or rotary dials on these early telephones, all requests from the customer to the "network" were by voice, e.g., "Can I speak with John?" or "Is the Doctor in today?" Over time, with the advent of rotary dial and ultimately dial-tone service, the ability to call "by name" or directly access services by voice requests was replaced by the necessity to dial directly to phone numbers (associated with people or services). Thus it became impossible to interact with humans for "simple" information requests such as "What movie is playing downtown?" The intelligent personalized services that were offered by operators like Mabel all but disappeared.

The challenge today is to build the twenty-first century version of Mabel—namely, an intelligent, personalized, communications agent, accessible across a range of pow-

erful, multimodal user interfaces (i.e., speech, text, graphics) that are available in a consistent way across a range of communications channels and devices. We call an early instantiation of such a system the *VoiceTone* system.

The promise of *VoiceTone* is to provide a powerful, yet natural, voice interface to all of the functionality and information available in the telecommunications network, the desktop, and the Internet. Whenever a user accesses the *VoiceTone* network, he or she is greeted with the generic spoken response of "How May I Help You?" The *VoiceTone* system provides integrated access to people, messages, news, directories, and other information services—all coordinated across network-based voice mail, e-mail, calendars, and address books.

A personalized and automated communications agent lies at the heart of the *VoiceTone* system. The "intelligent agent" will not only understand complex spoken requests (expressed in an unconstrained, conversational manner), but also be able to act on them. It will be able to learn and adapt to the customer's behavior and store the appropriate information about each customer within the network. Customers will be able to access a consistent set of service features, via consistent interfaces, by voice commands, using small hand-held devices, or from the desktop.

The first service features that will be part of the *VoiceTone* service will be voice dialing (using both a user-based and a common address book) and voice access to unified messaging services. Other potential *VoiceTone* services include the following.

1) Automated spoken access to personal and network directories, including white and yellow pages, and community directories, web searches, and buddy lists (e.g., "Connect me to Jane Doe in Memphis").
2) Screening of inbound communications, by automatically *identifying* voices and faces, and by querying callers (e.g., "John is calling you. Would you like to take the call?").
3) Conversion of media, so that text messages (e.g., e-mail) can be listened to, or voice messages read using a video display.
4) One-stop information access to any information service on the network. Hence, a user can access a weather service with the query "What's the weather in Boston?" or an airlines reservation service with the query "What flights from Newark to Seattle are available tomorrow?"
5) Searching and browsing spoken and typed message stores (i.e., voice mail and e-mail) in order to automatically prioritize and summarize messages (e.g., "Are there any messages from my wife?").
6) Managing calendars; for example, the *agent* can coordinate finding the best time for a meeting (including communicating with the personalized agents of other attendees).
7) Taking notes during conversations.
8) Language translation of messages, and, ultimately, live conversations.
9) Assisting with the communication needs of hearing-, speech-, and sight-impaired users.
10) Simple, yet powerful, access and interactivity with entertainment and broadcast news (e.g., "Suggest an old James Dean movie that I would like" or "Show me the Jets football game from the end zone camera").

The above list of functionality is not meant to be complete, but merely illustrative of voice-enabled services that are being investigated today.

*Beyond Voice Telephony—Multimodal Interactive Systems:* As small keyboardless PDAs with voice input and wireless telephony become more prevalent, the possibility arises that speech will be the input modality of choice in many situations, with gesture or pointing to the screen used as an alternative mode, primarily for selecting among a small set of options. Currently, applications have generally only considered whether to use speech input in lieu of DTMF input, and have not considered the simultaneous availability of multiple input modes (e.g., stylus and speech). Nor have they made use of any synergies that might be exploited when both visual and audio output modes are available. For example, in a directory query system, a speech-only system must rely on subdialogues to resolve ambiguity (e.g., in the multiple "John Lee" example above). If both screen and audio output are available, the dialogue could be altered to present the list of possibilities visually, along with the disambiguating fields, and simply ask the user to select the desired option. Research in multimodal interfaces is still in its infancy, with much of the work focusing on how to combine input to resolve ambiguity or conflict among input modes. Prototype multimodal systems that integrate speech and gesture have been demonstrated for resource deployment tasks using maps [78]. This work is aimed at defining a unification grammar for combining and reasoning about inputs from multiple modalities. Understanding how to use multiple modalities effectively, both for input and output, will be key to providing coherent and robust interactive services on next-generation access devices.

## V. Conclusion

With the rapid explosion of the web, and the associated communications services that are being offered to solve problems related to virtually every aspect of daily life, it is essential that such services be made available to as broad a population as possible. The only currently available ubiquitous access device to the network is the telephone, and the only ubiquitous user access technology mode is spoken voice commands and natural language dialogues with machines. In this paper, we have shown how all speech technologies have progressed to the point where they are now viable for a broad range of communications services, including compression of speech for use over wired and wireless networks, speech synthesis, recognition, and understanding for dialogue access to information, people, and messaging, and speaker verification for secure access to information and services. We discussed some of the unique

properties of wireless, POTS, and IP networks that make voice communication and control problematic. Finally, we discussed the types of voice services available in the past and today, and those that we foresee becoming available over the next several years.

## REFERENCES


[1] J. C. Ramming, "PML: A language interface to networked voice response units," in *Proc. Workshop on Internet Programming Languages*, Chicago, IL, May 1998.

[2] H. Dudley, "The vocoder," *Bell Lab. Rec.*, vol. 18, pp. 122–126, 1939.

[3] D. W. Petr, "32 kb/s ADPCM-DLQ coding for network applications," in *Proc. IEEE GLOBECOM*, Dec. 1982, pp. A8.3-1–A8.3-5.

[4] P. Kroon and W. B. Kleijn, "Linear-prediction based analysis-by-synthesis coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Asterdam, The Netherlands: Elsevier, 1995, pp. 79–119.

[5] R. E. Crochiere and J. M. Tribolet, "Frequency domain coding of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 512–530, 1979.

[6] A. Crossman, "A variable bit rate audio coder for videoconferencing," in *Proc. IEEE Workshop on Speech Coding for Telecommunications*, Sept. 1993, pp. 29–30.

[7] A. McCree, K. Truong, E. George, T. Barnwell, and V. Viswanathan, "A 2.4 kbit/s MELP coder candidate for the new US Federal Standard," in *Proc. ICASSP'96*, vol. 1, 1996, pp. 200–203.

[8] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Asterdam, The Netherlands: Elsevier, 1995, pp. 121–173.

[9] W. B. Kleijn and J. Haagen, "Waveform interpolation for coding and synthesis," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Asterdam, The Netherlands: Elsevier, 1995, pp. 175–207.

[10] *A Packet Loss Concealment Technique for Use with ITU-T Recommendation G.711*, ANSI Standard T1.521–1999, May 1999.

[11] J. M. Pickett, J. Schroeter, C. Bickley, A. Syrdal, and D. Kewley-Port, "Speech technology," in *The Acoustics of Speech Communication*, J. M. Picket, Ed. Boston, MA: Allyn and Bacon, 1998, ch. 17, pp. 324–342.

[12] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP'96*, 1996, pp. 373–376.

[13] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T next-gen TTS system," *J. Acoust. Soc. Amer.*, pt. 2, vol. 105, no. 2, p. 1030, 1999.

[14] J. Ostermann, "Animated talking head with personalized 3D Head Model," *J. VLSI Signal Process.*, vol. 20, pp. 97–105, 1998.

[15] E. Cosatto and H. P. Graf, "Sample-based synthesis of photo-realistic talking-heads," in *Proc. Computer Animation*, 1998, pp. 103–110.

[16] J. Ostermann, "Animation of synthetic faces in MPEG-4," in *Proc. Computer Animation*, 1998, pp. 49–55.

[17] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[18] J. Makhoul and J. Schwartz, "State of the art in continuous speech recognition," in *Voice Communication between Humans and Machines.*, D. Roe and J. Wilpon, Eds. Washington, DC: National Academy Press, 1994, pp. 165–188.

[19] G. Riccardi, R. Pieraccini, and E. Bocchieri, "Stochastic automata for language modeling," *Comput. Speech Lang.*, vol. 10, pp. 265–293, 1996.

[20] G. Riccardi and A. Gorin, "Stochastic language models for speech recognition and understanding," in *Proc. ICSLP 98*, 1998, pp. 2087–2090.

[21] R. Rose and E. Lleida, "Utterance verification in continuous speech recognition: Decoding and training procedures," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 126–139, Mar. 2000.

[22] M. Mohri and M. Riley, "Network optimization fo large-vocabulary speech recognition," *Speech Commun.*, vol. 28, pp. 1–12, 1999.

[23] M. Mohri, F. Pereira, and M. Riley, " Weighed automata in text and speech processing," presented at the *Proc. ECAI-96 Workshop*, Budapest, Hungary, 1996.

[24] M. Rahim, G. Riccardi, J. Wright, B. Buntschuh, and A. Gorin, "Robust automatic speech recognition in a natural spoken dialogue," presented at the *Proc. Workshop Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, 1999.

[25] D. S. Pallet, J. G. Fiscus, J. S. Garofolo, A. Martin, and M. A. Przybocki, "1998 Broadcast News Benchmark Test Results," presented at the *1999 DARPA Broadcast News Workshop*, Herndon, VA, 1999.

[26] A. Martin, J. Fiscus, M. Przybocki, and W. Fisher, "The evaluation: Word error rates and confidence analysis," presented at the *Proc. 9th Hub-5 Conversational Speech Recognition Workshop*, Linthicum Heights, MD, Sept. 24–25, 1998.

[27] D. Pallet, *et al.*, "1994 benchmark tests for the ARPA spoken language understanding program," in *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, Jan. 22–25, 1995, pp. 5–36.

[28] M. Mohri and M. Riley, "Integrated context-dependent networks in very large vocabulary speech recognition," in *Proc. 6th Euro. Conf. Speech Communication and Technology*, Budapest, Hungary, 1999, pp. 811–814.

[29] M. Mohri, M. Riley, D. Hindle, A. Ljolje, and F. Pereira, "Full expansion of context-dependent networks in large vocabulary speech recognition," presented at the *Proc. ICASSP'98*, vol. II, 1998, pp. 665–668.

[30] *Proc. DARPA Speech Recognition Workshop*, Harriman, NY, Feb. 18–21, 1996.

[31] A. Corazza and R. De Mori, "On the use of formal grammars," in *Spoken Dialogues with Computers*, R. De Mori, Ed. London, U.K.: Academic, 1998, pp. 461–484.

[32] R. Kuhn and R. De Mori, "Sentence interpretation," in *Spoken Dialogues with Computers*, R. De Mori, Ed. London, U.K.: Academic, 1998, pp. 486–522.

[33] B. C. Bruce, "Natural communication between person and computer," in *Strategies for Natural Language Processing*, W. G. Lehnert and M. H. Ringle, Eds. Hillsdale, NJ: Lawrence Erlbaum, 1982, pp. 55–88.

[34] M. Bates, "Models of natural language understanding," in *Voice Communication Between Humans and Machines*, D. Roe and J. Wilpon, Eds. Washington, DC: National Academy Press, 1994, pp. 238–253.

[35] L. Hirschman, "The roles of language processing in a spoken language interface," in *Voice Communication Between Humans and Machines*, D. Roe and J. Wilpon, Eds. Washington, DC: National Academy Press, 1994, pp. 217–237.

[36] S. Seneff, "Robust parsing for spoken language systems," in *Proc. ICASSP*, 1992, pp. 189–192.

[37] M. K. Brown and B. Buntschuh, "A context-free grammar compiler for speech understanding systems," in *Proc. ICSLP 94*, 1994, pp. 21–24.

[38] R. Pieraccini and E. Levin, "A spontaneous understanding system for database query applications," presented at the *Proc. ESCA Workshop on Spoken Dialogue Systems—Theories and Applications*, Vigso, Denmark, 1995.

[39] J. Wright, A. Gorin, and A. Abella, "Spoken language understanding within dialogs using a graphical model of task structure," presented at the *Proc. ICSLP 98*, 1998.

[40] A. L. Gorin, G. Riccardi, and J. H. Wright, "How may I help you?," *Speech Commun.*, vol. 23, pp. 113–127, 1997.

[41] A. Gorin, S. Levinson, L. Miller, A. Gertner, E. Goldman, and A. Ljolje, "On adaptive acquisition of language," in *Proc. ICASSP'90*, 1990, pp. 601–604.

[42] A. L. Gorin, "On automated language acquisition," *J. Acoust. Soc. Amer.*, vol. 97, pp. 3441–3461, 1995.

[43] M. Denecke and A. Waibel, "Dialogue strategies guiding users to their communicative goals," in *Proc. Eurospeech 1997*, 1997, pp. 1339–1342.

[44] G. Riccardi and A. L. Gorin, "Stochastic language adaptation over time and state in natural spoken dialogue systems," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 3–10, Jan. 2000.

[45] C. Sorin and R. De Mori, "Sentence generation," in *Spoken Dialogues with Computers*, R. De Mori, Ed. London, U.K.: Academic, 1998, pp. 563–582.

[46] C. A. McCann, W. Edmonson, and R. K. Moore, "Practical issues in dialogue design," in *The Structure of Multimodal Dialogue*, M. M. Taylor, F. Neel, and D. G. Bouwhuis, Eds. Amsterdam, The Netherlands: Elsevier, 1989, pp. 467–480.

[47] E. Levin, R. Pieraccini, and W. Eckert, "Learning dialogue strategies within the Markov decision process framework," in *Proc. 1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997, pp. 72–79.
[48] M. Walker, D. Hindle, J. Fromer, G. DiFabbrizio, and C. Mestel, "Evaluating competing agent strategies for a voice email agent," presented at the *Proc. 5th Euro. Conf. Speech Communication and Technology (EUROSPEECH97)*, 1997.
[49] R. Billi, G. Castagneri, and M. Danieli, "Field trial evaluations of two different information inquiry systems," in *Proc. IVTTA*, 1996, pp. 129–134.
[50] D. J. Litman and S. Pan, "Empirically evaluating an adaptable spoken dialogue system," presented at the *Proc. 7th Int. Conf. User Modeling*, 1999.
[51] S. Carberry, J. Chu-Carroll, and L. Lambert, "Modeling intention: Issues for spoken language dialogue systems," in *Proc. 1996 Int. Symp. Spoken Dialogue, ISSD 96*, 1996, pp. 13–24.
[52] M. Walker, "A peak, a plateau, or a stiff climb?," *ELSNEWS 8.1*, pp. 1–3, Winter 1999.
[53] S. E. Brennan, "The grounding problem in conversation with and through computers," in *Social and Cognitive Psychological Approaches to Interpersonal Communication*, S. R. Fussell and R. J. Keutz, Eds. Hillsdale, NJ: Lawrence Erlbaum, 1998, pp. 201–255.
[54] M. Walker, D. Litman, C. Kamm, and A. Abella, "PARADISE: A general framework for evaluation spoken dialogue agents," in *Proc. 35th Annu. Meeting of the Association of Computational Linguistics, ACL/EACL 97*, 1997, pp. 271–280.
[55] C. Kamm, M. Walker, and D. Litman, "Evaluating spoken language systems," in *Proc. AVIOS 1999*, 1999, pp. 187–197.
[56] K. Yu and J. Mason, "On-line incremental adaptation for speaker verification using maximum likelihood estimates of CDHMM parameters," in *Proc. ICSLP'96*, 1996, pp. 1752–1755.
[57] A. E. Rosenberg and S. Parthasarathy, "Speaker identification with user-selected password phrases," in *Proc. Eurospeech'97*, 1997, pp. 1371–1374.
[58] T. Nordstrom, H. Melin, and J. Lindberg, "A comparative study of speaker verification systems using the Polycast database," in *Proc. ICSLP'98*, 1998, pp. 1359–1362.
[59] Y. Gu and T. Thomas, "An implementation and evaluation of an on-line speaker verification system for field trials," in *Proc. ICSLP'98*, 1998, pp. 125–128.
[60] J. Choi, D. Hindle, J. Hirschberg, I. Magrin-Chagnolleau, C. Nakatani, F. Pereira, A. Singhal, and S. Whittaker, "SCAN—Speech content based audio navigator: A systems overview," in *Proc. ICSLP98*, 1998, pp. 2861–2864.
[61] R. V. Cox, B. Haskell, Y. LeCun, B. Shahraray, and L. Rabiner, "On the application of multimedia processing to telecommunications," *Proc. IEEE*, vol. 86, pp. 755–824, May 1998.
[62] Q. Huang, Z. Liu, A. Rosenberg, D. Gibbon, and B. Shahraray, "Automated generation of news content hierarchy by integrating audio, video, and text information," in *Proc. ICASSP'99*, vol. VI, 1999, pp. 3025–3028.
[63] Z. Liu and Q. Huang, "Classification of audio events for broadcast news," in *IEEE Workshop Multimedia Signal Processing*, Los Angeles, CA, Dec. 1998, pp. 364–369.
[64] R. Nakatsu, "ANSER: An application of speech technology to the Japanese banking industry," *Computer*, vol. 23, pp. 43–48, Aug. 1990.
[65] S. Furui, private communication.
[66] M. Lennig, "Putting speech recognition to work in the telephone network," *Computer*, vol. 23, pp. 35–41, Aug. 1990.
[67] J. G. Wilpon, L. Rabiner, C.-H. Lee, and E. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 1870–1878, Nov. 1990.
[68] M. Yuschik, E. Schwab, and L. Griffith, "ACNA—The Ameritech customer name and address servicer," *J. AVIOS*, vol. 15, pp. 21–33, 1994.
[69] D. Yashchin, S. Basson, A. Kalyanswamy, and K. Silverman, "Results from automating a name & address service with speech synthesis," in *Proc. AVIOS*, 1992.
[70] A. Sanderman, E. den Os, A. Cremers, L. Boves, and J. Sturm, "Evaluation of the dutch train timetable information system developed in the ARISE project," in *Proc. IVTTA 98*, 1998, pp. 91–96.
[71] G. Castagneri, P. Baggia, and M. Danieli, "Field trials of the Italian ARISE train timetable system," in *Proc. IVTTA 98*, 1998, pp. 97–102.
[72] L. Lamel, S. Rossett, J. L. Gauvain, S. Bennacef, M. Garnier-Rizet, and B. Prouts, "The LIMSI ARISE system," in *Proc. IVTTA 98*, 1998, pp. 209–214.
[73] S. M. Marcus, D. W. Brown, R. G. Goldberg, M. S. Schoeffler, W. R. Wetzel, and R. R. Rosinski, "Prompt constrained natural language—Evolving the next generation of telephony services," in *Proc. ISSD 96*, 1996.
[74] M. Lennig, G. Bielby, and J. Massicote, "Directory assistance automation in Bell Canada: Trial results," in *Proc. IVTTA 94*, 1994, pp. 8–12.
[75] V. Gupta, S. Robillard, and C. Pelletier, "Automation of locality recognition in ADAS Plus," in *Proc. IVTTA 98*, 1998, pp. 1–4.
[76] R. Billi, F. Canavesio, and C. Rullent, "Automation of telecom Italia directory assistance service: Field trial results," in *Proc. IVTTA 98*, 1998, pp. 11–14.
[77] B. Buntschuh, C. Kamm, G. DiFabbrizio, A. Abella, M. Mohri, S. Narayanan, I. Zeljkovic, R. D. Sharp, J. Wright, J. Shaffer, R. Duncan, and J. Wilpon, "VPQ: A spoken language interface to large scale directory information," presented at the *Proc. ICSLP 98*, 1998.
[78] M. Johnston, P. R. Cohen, D. McGee, S. L. Oviatt, J. A. Pittman, and I. Smith, "Unification-based multimodal integration," presented at the *Proc. 35th Annu. Meeting Association of Computational Linguistics*, ACL/EACL, 1997.
[79] H. Dudley, "The vocoder—Electrical re-creation of speech," *J. Soc. Motion Pict. Eng.*, vol. 34, pp. 272–278, 1940.
[80] *Appendix I—A high quality low-complexity algorithm for packet loss concealment with G.711*, ITU-T Recommendation G.711, May 1999.
[81] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal. The AT&T next-gen TTS system. presented at *Proc. Joint Meeting of ASA, EAA, and DEGA*. [CD-ROM] Available: http://www.research.att.com/projects/tts

**Richard V. Cox** (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ.

In 1979, he joined the Acoustics Research Department, Bell Laboratories. He conducted research in the areas of speech coding, digital signal processing, analog voice privacy, audio coding, and real-time implementations. In 1987, he was promoted to Supervisor of the Digital Principles Research Group. The group implemented sophisticated digital compression algorithms for speech, audio, image, and video coding. During this period, he became engaged in speech coding standards work for digital cellular telephony and within the International Telecommunications Union. He collaborated on the low-delay CELP algorithm that became ITU-T Recommendation G.728 in 1992. He managed the ITU effort that resulted in the creation of ITU-T Rec. G.723.1 in 1995. In 1992, he was appointed Department Head of the Speech Coding Research Department of AT&T Bell Laboratories. In 1996, as part of the split-up of AT&T, he joined AT&T Labs-Research, Florham Park, NJ, as Division Manager of the Speech Processing Software and Technology Research Department. In 2000, he was appointed Director of the Speech and Image Processing Services Research Laboratory.

Dr. Cox presently serves as President-Elect of the IEEE Signal Processing Society. He is also a member of the Board of Directors of Recording for the Blind & Dyslexic, the only U.S. provider of textbooks and reference books for people with print disabilities. At RFB&D, he is presently helping to lead the effort to bring out digital books combining audio, text, images, and graphics for their consumers.

**Candace A. Kamm** (Senior Member, IEEE) received the B.S. degree in psychology from the University of California, Santa Barbara, the M.S. degree in audiology from California State University, Los Angeles, and the Ph.D. degree in cognitive psychology from the University of California, Los Angeles (UCLA).

She is Division Manager of the Human–Computer Interaction Research Department at the Information Systems and Services Research Laboratory, AT&T Labs-Research, Florham Park, NJ. She was a Research Audiologist at UCLA School of Medicine from 1974 to 1982, where her work focused on speech recognition performance by normal-hearing and hearing-impaired listeners, evaluation of reliability of speech recognition tests, speech recognition in noise, and application of adaptive procedures for measurement of maximum speech recognition. In 1982, she became a Member of Technical Staff at Bell Laboratories, evaluating performance of automatic speech recognition systems. At the divestiture of the Bell System in 1984, she joined Bellcore, were she worked on various aspects of speech recognition and speech synthesis technology and applications, and became Director of the Speech Recognition Applications Research Group in 1993. In 1995, she returned to AT&T Bell Laboratories, and has been a member of AT&T Labs-Research since its inception. Her recent research interests include design and evaluation of spoken-dialogue systems and multimodal speech-enabled systems.

Dr. Kamm is a Member of the IEEE Signal Processing Society, currently serving on the Board of Governors, and a member of the Association for Computing Machinery, the Association for Computational Linguistics, and the Acoustical Society of America.

**Lawrence R. Rabiner** (Fellow, IEEE) was born in Brooklyn, NY, on September 28, 1943. He received the S.B. and S.M. degrees simultaneously in 1964, and the Ph.D. degree in electrical engineering in 1967, all from the Massachusetts Institute of Technology, Cambridge, MA.

From 1962 to 1964, he participated in the cooperative program in Electrical Engineering at AT&T Bell Laboratories, Whippany and Murray Hill, NJ. During this period, he worked on digital circuitry, military communications problems, and problems in binaural hearing. He joined AT&T Bell Laboratories in 1967 as a Member of Technical Staff. He was promoted to Supervisor in 1972, Department Head in 1985, Director in 1990, and Functional Vice President in 1995. He joined the newly created AT&T Labs-Research, Florham Park, NJ, in 1996 as Director of the Speech and Image Processing Services Research Laboratory, and was promoted to Vice President of Research in 1998. In his current role, he manages a broad research program in communications, computing, and information sciences technologies. He is coauthor of the books *Theory and Application of Digital Signal Processing* (Englewood Cliffs, NJ: Prentice-Hall, 1975), *Digital Processing of Speech Signals* (Englewood Cliffs, NJ: Prentice-Hall, 1978), *Multirate Digital Signal Processing* (Englewood Cliffs, NJ: Prentice-Hall, 1983), and *Fundamentals of Speech Recognition* (Englewood Cliffs, NJ: Prentice-Hall, 1993).

Dr. Rabiner is a member of Eta Kappa Nu, Sigma Xi, Tau Beta Pi, the National Academy of Engineering, the National Academy of Sciences, and a Fellow of the Acoustical Society of America, Bell Laboratories, and AT&T. He is a former President of the IEEE Acoustics, Speech, and Signal Processing Society, a former Vice-President of the Acoustical Society of America, a former editor of the TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, and a former member of the PROCEEDINGS OF THE IEEE Editorial Board.

**Juergen Schroeter** (Senior Member, IEEE) received the Dipl.-Ing. degree (EE) in 1976, and the Dr.-Ing. degree (EE) in 1983 from Ruhr-Universitaet Bochum, Germany.

From 1976 to 1985, he was with the Institute for Communication Acoustics, Ruhr-University Bochum. From 1986 to 1995, he was with AT&T Bell Laboratories, where he worked on speech coding and synthesis methods employing models of the vocal tract and vocal cords. In 1996, he joined AT&T Labs-Research, Florham Park, NJ, as a Division Manager. He is currently leading the team that created AT&T's Next-Generation Text-to-Speech system.

Dr. Schroeter is a Fellow of ASA.

**Jay G. Wilpon** (Fellow, IEEE) is Division Manager of Speech Processing Software and Technology Research within AT&T Labs-Research, Florham Park, NJ. Since joining AT&T in June 1977, he has concentrated his research on problems in automatic speech recognition. He has more than 90 publications and has been awarded nine patents. He is the co-editor of the book *Voice Communication Between Humans and Machines*, (Washington, DC: National Academy of Sciences-Nat. Res. Council). His current research interests include several of the key problems that will promote the ubiquitous use of speech recognition technology within the communications industry. The focus of this work is on robust speech recognition algorithms, and spoken language understanding and dialogue management with emphasis on their application to support new innovative user interfaces to services.

Mr. Wilpon served as Member at Large to the IEEE SPS Board of Governors from 1996 to 1999 and as Chair of the IEEE Signal Processing Society's (SPS) Speech Processing Technical Committee from 1993 to 1995. In 1987, he received the IEEE Acoustics, Speech, and Signal Processing Society's Paper Award for his work on clustering algorithms used in training automatic speech recognition systems.