Voice Access of Global Information for Broad-Band Wireless: Technologies of Today and Challenges of Tomorrow

LIN-SHAN LEE, FELLOW, IEEE, AND YUMIN LEE, MEMBER, IEEE

Invited Paper

The rapid development of the Internet and the World Wide Web has created a global network that will soon become a physical embodiment of the entire human knowledge and a complete integration of the global information activities. The traditional approach to access the network is through a computer physically tied to the network. As broad-band wireless takes off, the traditional tethered approach will gradually become obsolete. It is believed that one of the most natural and user-friendly approaches for accessing the network will be via human voice, and the integration of spoke language processing technologies with broad-band wireless technologies will be a key to the evolution of a broad-band wireless information community. This paper offers a vision of the above concept. Technical considerations and some typical example applications of accessing the information and services using voice in a broad-band wireless environment are discussed. Fundamentals of spoken language processing technologies that are crucial in such a broad-band wireless environment are briefly reviewed. Technical challenges caused by the unique nature of wireless mobile communications are also presented along with some possible solutions.

Keywords*—*Information retrieval, radio communication, speech processing.

I. INTRODUCTION

People were dreaming of a universally accessible information database even before computer networking was invented. It was not until the 1990s that the technologies caught up to make such a concept possible. The commercialization of the Internet and the invention of the World Wide Web have prompted the creation of a universe of virtually unlimited, network-accessible information that ranges from private, secure databases to free, publicly available services. Such typical examples as digital libraries, virtual museums, distance learning, network entertainment, network banking, and electronic commerce have indicated that this global information network will soon become a physical embodiment of the whole human knowledge, and a complete integration of the global information activities. As the global information network evolves, efficient, while user friendly, technologies for accessing the information infrastructure become increasingly important. Traditionally, the information infrastructure is accessed through a computer that is physically tied to a network. Commands are entered into the computer by the use of the keyboard and the mouse, while the retrieved information is displayed on the screen. Although reliable and economical, this mode of accessing the global information network will become increasingly inadequate as we progress into the 21st century.

The success of broad-band wireless technologies will bring a new dimension to the way the above information is accessed or retrieved. Since full mobility support is possible, the user will no longer be constrained by the tether of the wire. The data rates of the future broad-band wireless systems range from 384 kb/s [third-generation (3G) cellular systems] to a few megabits per second (broad-band wireless access) to tens of megabits per second (wireless LAN), which are capable of carrying multimedia traffic. Therefore, with the successful deployment of broad-band wireless communication systems, users should be able in principle to access the global network infrastructure at any time and from anywhere. However, as the size of the wireless terminal shrinks, traditional keyboard or keypad input and screen output become increasingly inconvenient. Therefore, an efficient, flexible, and user-friendly approach especially suited for broad-band personal wireless communicators is an important enabling factor for desired "anytime, anywhere" access to the information infrastructure. Speech is one of the most natural and user-friendly mechanisms for informa-

Manuscript received February 13, 2000; revised September 3, 2000.

The authors are with the Department of Electrical Engineering and Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan 10617, R.O.C.

Publisher Item Identifier S 0018-9219(01)00451-0.

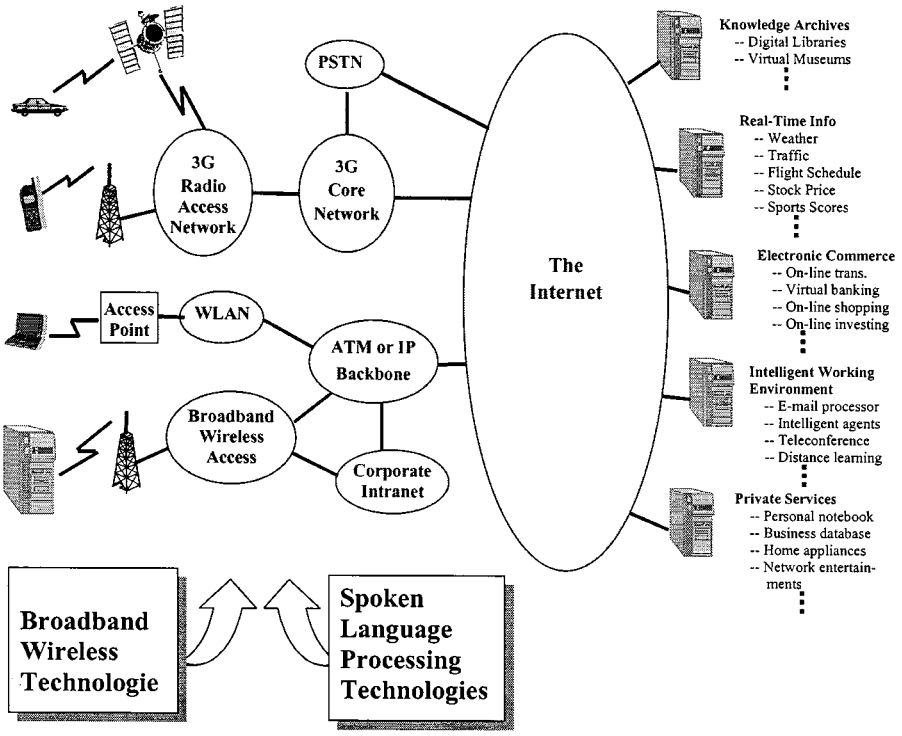

Fig. 1. The broad-band wireless environment in which the mobile users have ubiquitous tetherless broad-band access to the global information network. The integration of spoken language processing and broad-band wireless technologies offers the future voice-enabled information community.

tion access, and spoken language processing technologies, after decades of diligent research efforts, have matured to a point where many useful and commercially beneficial applications have recently become feasible. Therefore, a good integration of advanced spoken language processing and broad-band wireless technologies will be a key factor for the evolution of a wireless information community from the success of broad-band wireless. Such a concept is depicted in Fig. 1.

In this paper, we offer our vision of the integration of spoken language processing technologies with broad-band wireless communications technologies. Emerging broad-band wireless technologies are briefly summarized in Section II. Example applications, including voice-enabled information retrieval, remote authoring, and interactive voice access to personalized intelligent agents, and technical considerations of accessing the information and services using voice in the broad-band wireless environment are discussed in detail in Section III. Fundamentals of spoken language processing technologies and speech processing functionalities that are crucial in the future broad-band wireless network environment are presented in Sections IV and V. Technical challenges that arise due to the unique nature of wireless and mobile communications, including poor microphone reception, background noise, transmission and packet loss, bandwidth limitations, hardware and size limitations for handsets, and cross-language information processing necessary to support global roaming are also discussed in these sections. Finally, the conclusion is given in Section VI.

II. The Broad-Band Wireless Environment

Wireless communication networks have traditionally been used primarily to carry circuit-switched voices. However, due to the phenomenal growth of the Internet and the ever more abundant information and services that are available on the web, users increasingly desire the ability to access the global information network using their wireless terminal at any time, from anywhere, and with the same reliability and speed provided by fixed networks. This demand for a reliable, high-speed, and ubiquitous wireless access network has stimulated substantial efforts of research and standardization on broad-band wireless technologies.

In essence, broad-band wireless refers to the multimedia-capable high-data-rate (greater than 384 kb/s) services provided to mobile wireless terminals. The ultimate goal of broad-band wireless is to provide services to mobile users with the same high quality and broad-band characteristics as fixed networks without the tether of the wire. Mobility should be fully and seamlessly supported, i.e., users should receive exactly the same services wherever and whenever they access the network. Physical limitations and channel impairments, including limited bandwidth, power and complexity constraints, multipath propagation, signal fading, interference, and noise, present significant technical hurdles that must be overcome in order to provide reliable high-data-rate wireless mobile communication services. The efforts to overcome these technical challenges have led to several emerging system approaches, including the evolution of second-generation cellular systems to higher data rates [1]–[8], the so-called 3G cellular system [8]–[19],

broad-band wireless local access systems, and wireless local area networks (WLANs), all of which are briefly summarized later in this section. Although these emerging systems sometimes compete with each other, it has become a common understanding that different broad-band wireless technologies must peacefully coexist and somewhat interoperate with each other in order to achieve the goal of providing ubiquitous tetherless broad-band access to the global information network. Flexibility, globalization, harmonization, and convergence are therefore important trends in the development of these broad-band wireless technologies. Thus, although the emerging technologies may differ in many aspects such as air interface, data rate, coverage area, and supported mobile speed, interoperability, and interworking together with advanced technologies such as software-definable radio [20], [21] are expected to enable seamless information transfer among different systems in the future. As shown in Fig. 1, despite the technological differences, the emerging systems promise the creation of a future "broad-band wireless environment" in which the mobile users have ubiquitous tetherless broad-band access to the global information network. A good integration of advanced spoken language processing techniques into this broad-band wireless environment will be a key factor in the evolution of a wireless information community, as discussed in this paper.

A. Evolution of Second-Generation Cellular Systems to Higher Data Rates

These systems extend the capabilities of existing services and systems to provide an orderly and planned transition into broad-band wireless. Examples include EDGE [Enhanced Data Rates for Global System for Mobile (GSM) Evolution] [1], [2], UWC-136 [3]–[5], [8], and CDMA2000 [6], [7]. EDGE is a high-data-rate extension of GSM. By using eight-level phase shift keying (8-PSK) in 200-kHz carriers and advanced link quality control schemes, EDGE is capable of providing data rates in excess of 384 kp/s in outdoor vehicular environments [1], [2]. On the other hand, UWC-136 has been adopted by the Universal Wireless Communications Consortium (UWCC) for evolving the Telecommunications Industry Association (TIA) Interim Standard 136 (IS-136) time-division multiple-access (TDMA)-based technology [3]–[5], and is capable of providing data rates up to 384 kb/s and 2 Mb/s in the outdoor and indoor environments, respectively. CDMA2000 is an evolution from the TIA IS-95 code-division multiple-access (CDMA) standard, and supports data rates up to 2 Mb/s [6], [7].

B. 3G Cellular System

The 3G cellular system refers to the International Mobile Telecommunication-2000 (IMT-2000) currently being standardized in the International Telecommunication Union (ITU). Formerly known as Future Public Land Mobile Telecommunication Systems (FPLMTS), the aim of IMT-2000 is to define a family of radio interfaces suitable for the wide range of radio operating environments, e.g., indoor, outdoor, terrestrial, and satellite, with target data rates of 384 kb/s for wide area coverage and 2 Mb/s for local area (microcellular) coverage. IMT-2000 is a global standardization effort with participation from Europe [8], [12], Asia-Pacific [13]–[16], and the United States [17], [18]. A comprehensive set of terrestrial and satellite radio interface specifications were approved in November 1999 [19]. The approved terrestrial radio interfaces employ a wide range of radio technologies that are newly designed for accessing existing core networks (e.g., UTRA [8], [12]) or evolved from the second-generation cellular systems (e.g., CMDA2000 [17] and UWC-136 [18]). The approved satellite interfaces cover LEO, MEO, and GEO orbits as well as those specifically aimed at maximizing the commonality between terrestrial and satellite interfaces [19].

C. Broad-Band Wireless Local Access Systems

Broad-band wireless local access systems are wireless extensions (typically with data rates in excess of 2 Mb/s) of fixed broad-band networks, which provide the users with a means of radio access to broad-band services. Research and development activities worldwide [22]–[26] have led to emerging systems that differ in applications, coverage area, data rate, and scale of mobility. For example, Magic Wireless ATM Network Demonstrator (Magic WAND) [25] and ATM Wireless Access Communication System (AWACS) [25] are experimental low-mobility indoor wireless systems for accessing the ATM infrastructure. High-Performance Radio Local Area Network (HIPERLAN) Type 2 [27] and HIPERACCESS [28], on the other hand, are members of the Broadband Radio Access Networks (BRAN) family of standards currently being developed within the European Telecommunication Standards Institute (ETSI). Both of these standards are aimed at providing low-mobility users wireless access of IP, ATM, and possibly Universal Mobile Telecommunication System (UMTS) core networks at a data rate up to 54 Mb/s. HIPERLAN Type 2 is designed primarily for indoor applications, while HIPERACCESS is designed for remote access applications. The local multipoint distribution services (LMDS) and multipoint multichannel distribution services (MMDS) [29] are example approaches for delivering broad-band wireless access to residential and commercial customers, and are being considered for standardization by the IEEE 802.16 Working Group [30]. Finally, System for Advanced Mobile Broadband Applications (SAMBA) is a trial platform for demonstrating full-duplex wireless ATM access with high-mobility (50 km/h) support [25].

D. WLANs

WLANs [31] are high-speed (data rate in excess of 1 Mb/s) indoor radio networks with limited support for mobility. Existing WLANs typically support two network topologies: a "centralized" topology in which the mobile terminals access the fixed backbone network via access points, and a "distributed" topology in which a group of mobile terminals communicate with each other in an ad hoc fashion. Therefore, WLANs can also be viewed as a special case of the previously mentioned broad-band wireless local access sys-

tems. Examples of WLAN standards include IEEE 802.11 Wireless LAN [32] and HIPERLAN Type 1 [33], [34]. IEEE 802.11 Wireless LAN includes both radio and infrared (IR) light implementations that provide data rates from 1 to up to 54 Mb/s [27]. HIPERLAN Type 1 is a member of the BRAN family of standards, and is capable of delivering a channel data rate of 23.5 Mb/s.

III. Voice Access of Information in the Broadband Wireless Environment—Challenges of Tomorrow

Given that services such as wireless stock quotes and wireless sports score updates are already becoming popular today even with the limited data rates of the paging networks, it is not difficult to envision that the success of broad-band wireless will prompt the creation of many new applications that rely on the "anytime, anywhere" access to the information and services on the Internet or web. Therefore, future wireless terminals must provide users with a user-friendly interface for accessing the information on the network infrastructure at any time and from anywhere. Traditionally, the information on the network infrastructure is accessed through a computer physically tied to a network, with commands entered using the keyboard and the mouse and the retrieved information displayed on the monitor. However, in the broad-band wireless environment, user terminals are becoming increasingly miniaturized for better portability, rendering the traditional keyboard or keypad input and screen output practically inconvenient. Furthermore, with ever-cheaper digital and radio frequency (RF) technologies, many emerging broad-band wireless applications will require the embedding of digital and RF components in otherwise traditional appliances such as televisions, VCRs, refrigerators, microwave ovens, and car electronics. For these appliances, attaching a keyboard and a mouse is difficult and awkward, if not impossible. Finally, in order to fully exploit the ubiquitous access to the information infrastructure, the use of wireless terminals should not be limited to situations where the traditional information access method is viable. As a result, although the traditional mode of information access is reliable and economical, it will become increasingly inadequate as the new broad-band wireless applications take off. An efficient, flexible, and user-friendly information access mechanism suitable for broad-band wireless terminals is an important enabling factor for "anytime, anywhere" access to the information infrastructure.

One of the most convenient, user-friendly, and natural mechanisms for information access is via the human voice; therefore, it is not surprising that communication services with a voice-enabled access interface based on spoken language processing technologies date back as far as the early 1980s [35]. We further argue in this paper that voice access to the global information network will, in fact, be ideal in the future broad-band wireless environment for the following reasons. First, since speech is a natural part of everyone's daily life, users do not need any special training for using the voice access. Second, the only additional devices required for voice access are microphones and speakers, which are small and inconspicuous, and can be easily integrated with traditional or future wireless information appliances. In fact, many existing wireless or mobile terminals, such as cellular phones and notebook computers, already have built-in microphones and speakers. The only missing piece in these cases is the ability to process speech signals for information accessing applications, which can actually be provided primarily at the servers in the future client–server networks. Finally, voice access can be used almost anywhere and in almost any situation, and is, in fact, perfect for many "hands-busy, eyes-busy" scenarios such as automobile driving. Therefore, the omnipresence of the broad-band wireless, complemented by the convenience of voice access, shall offer users access to the global information network even in situations where traditional access mechanisms are impossible due to limitations in space (e.g., in a crowded subway), safety concerns (e.g., when driving an automobile), or the wireless terminal itself (e.g., for a pocket personal communicator). As a result, it is believed that a good integration of the advanced spoken language processing and broad-band wireless technologies is indeed a key factor for the evolution of a wireless information community from the success of broad-band wireless.

Possible activities of voice access of the global information network in the broad-band wireless environment can be roughly classified into three categories: voice-enabled information retrieval, remote authoring, and interactive voice access to personalized intelligent agents. Each of these categories presents a set of challenges and technology requirements, as briefly discussed below.

A. Voice-Enabled Information Retrieval

Voice-enabled information retrieval refers to the retrieval of information, using speech input and audio output (possibly complemented with visual output), from private/secure or public databases that are located on the information infrastructure. Perhaps the most obvious application is the voice browsing of the web. The World Wide Web Consortium (W3C), recognizing the importance of voice interface for web browsing, established a Voice Browser Activity and Work Group in March 1999 [36]. The success of voice browsing depends on voice-friendly rendering of the network contents as well as the integration of spoken language processing technologies into the browser technologies to create "voice browsers." Currently, there are generally two approaches for voice-friendly rendering of contents. One is to extend the Hyper Text Markup Language (HTML) using style sheets such as the Aural Cascading Style Sheets (ACSS) [37], which allows a document to be displayed aurally as well as visually without requiring a separate page for each mode. The other approach is to create a specific markup language for rendering speech input/output on the Internet. An example of the latter is the VoxML Voice Markup Language [38]. Current voice browser technologies include self voicing browser for the visually challenged [39], speech interface for accessing selected web-based databases [40], [41], telephone-based web browsers [42]–[44], voice

access of information for automobile drivers [45], [46], and voice portal technologies. Most of these examples use circuit-switched plain old telephone system (POTS) as an access device, and none of them employ speaker verification as a means for user authentication. Integrating the advances in broad-band wireless and spoken language processing technologies, future voice-enabled information retrieval can be significantly enhanced to include many more applications not available today. Typical examples may be voice access to a confidential patient database from an ambulance, and voice access to a secure inventory database from a delivery vehicle.

B. Remote Authoring

Remote authoring is yet another category of examples within the broad spectrum of applications that will be enabled by the integration of broad-band wireless and spoken language processing technologies. Consider the scenario in which a news reporter authors his story by dictating the article into a handheld wireless device, finding relevant multimedia information over the network, and composing the complete document via the same device. With current technology, the reporter will have to manually transcribe his article and find relevant information when he returns to the office. With remote authoring, he can find relevant multimedia information in real time, while the dictated article can be automatically converted into text ready for editing before he returns to the office. Other example scenarios include remote transcription of meetings, presentations, and interviews. Note that here remote authoring is different from voice-enabled information retrieval in that in remote authoring, the user actively creates and manipulates the information.

Two ingredients are necessary for the remote authoring to be feasible and practical. The first is the ability for the mobile user (e.g., reporter) to reliably access the network at any time and from anywhere, which is adequately provided by broad-band wireless. The second is a large-vocabulary dictation system that also supports mobility. Since a large-vocabulary dictation system often requires much more memory and processing power than a handheld device can handle, distributed speech recognition (DSR) is a more practical approach. DSR is a client/server approach for speech recognition, where the client refers to the mobile terminal, while the server is a computer physically connected to the network infrastructure. The mobile client may perform only the feature parameters extraction and compression. The compressed feature parameters are then transmitted over an error-protected wireless channel to the server, which is in charge of more complicated tasks of large-vocabulary speech recognition. The feature extraction, compression algorithms as well as the error correction code are important factors in the design of a DSR system. A well-designed DSR system strikes a good balance among mobile terminal complexity, wireless transmission bandwidth requirement, and speech recognition accuracy or information retrieval efficiency. DSR is actually useful for all possible applications of voice access of networked information in the broad-band wireless environment in general, and especially helpful to the remote authoring discussed here in particular, because a large-vocabulary speech recognition system that supports mobility is needed here. The Aurora project within ETSI is an ongoing effort to standardize feature extraction, compression, and error correction coding algorithms for DSR.

C. Interactive Voice Access to Personalized Intelligent Agents

Personalized intelligent agents, or personal assistants, refer to integrated information- and communication-related services that are customized for each user. The concept of integrated services is not new. In fact, the popularity of the Internet and the web has already prompted the appearance of web- and telephone-accessible services integrating voice dialing, personal address and appointment notebook, message retrieval (voice mail and e-mail), and information retrieval [47]. A futuristic voice-enabled interactive personalized intelligent agent referred to as *VoiceTone* may be able to provide secure and integrated access to messages, news, directories, personal information, and other information services [47]. In the broad-band wireless environment, an important and special requirement for such services is the ability to handle multilingual tasks caused by global roaming. Such features as language or dialect detection for the input speech signal plus machine translation may be required to provide users who travel across many different countries with the same high-quality and friendly voice interfaces.

D. Challenges of Tomorrow

Most existing spoken language processing applications, such as voice access to credit card accounts or directory assistance, provide voice access to information and services through the circuit-switched POTS. Future applications in the broad-band wireless environment must be made available through both the circuit-switched POTS and the packet-switched broad-band wireless networks using miniaturized portable terminals. This shift in paradigm presents many unique and significant challenges to spoken language processing technologies due to the nature of wireless communications.

Wireless channels are relatively unreliable due to multipath fading propagation. Furthermore, mobile wireless terminals are often used in noisy environments such as in a car or in a crowded shopping mall. The computational power of a mobile wireless terminal is often limited due to concerns of portability and battery life. These constraints bring about many technical challenges for voice access technologies applied to broad-band wireless. For example, the complexity and power drain of the speech processors built on the mobile wireless terminals must be low enough for the terminals to handle. Speech recognition technologies, if implemented on the mobile terminal, must deal with such impairments as variable and generally poor microphone reception and background noise. In cases where the speech signals or speech feature parameters are transmitted wirelessly to a speech recognition server that is located somewhere in the network,

Table 1
Basic Spoken Language Processing Technologies and Fundamental Speech Processing Functionalities Necessary for Voice Access of Information in the Broad-Band Wireless Environment

Technologies and Fuctionalities \ Applications		Voice-Enabled Information Retrieval	Remote Authoring	Voice Access to Intelligent Agents
Technologies	Speech Recognition	●	●	●
	Speaker Verifcation	●	●	●
	Speech Understanding	●	●	●
	Text-To-Speech Synthesis	●	●	●
Functionalities	Dictation and Transcription		●	●
	Audio Indexing and Retrieval	●		●
	Spoken Dialogue	●		●
	Multilingual Functionality	●	●	●

the speech signals or feature parameters must be tailored to fit in the limited bandwidth even when broad-band wireless is used. Speech recognition in these cases must also deal with such impairments as lost frames due to transmission errors or network congestion, low-bit-rate speech coding, as well as poor microphone reception and background noise, all of which result in significant speech quality degradation. Furthermore, in general speech recognition is required to be very robust and reliable because backup methods such as keypad or touch-tone input may become too cumbersome or may even no longer be available. Speaker verification technologies for user authentication also face similar challenges. On the other hand, text-to-speech synthesis and dialogue systems must deal with lost frames, variable and generally longer delay encountered in broad-band packet-switched networks, as well as degraded speech quality due to the limited bandwidth of wireless communications. Furthermore, when the users travel across many different countries with different languages, the global roaming may further produce multilingual difficulties. For example, the input speech may be in a language different from the ones that the local service providers primarily handle. Language identification and machine translation technologies will be needed in such cases to identify the user's language and perform the necessary translation.

IV. Enabling Spoken Language Processing Technologies

Voice access of global information in the broad-band wireless environment relies on the availability and maturity of enabling spoken language processing technologies and speech processing functionalities that are properly adapted and enhanced to meet the unique challenges presented by broad-band wireless communications. The speech processing functionalities include dictation and transcription, audio indexing and retrieval, spoken dialogue systems, and multilingual functionality, all of which will be discussed in Section V. Lying at the core of these functionalities are the enabling spoken language processing technologies, including large-vocabulary continuous speech recognition, speech understanding, speaker verification, and text-to-speech synthesis, all of which are summarized below. As shown in Table 1, the enabling spoken language processing technologies are required for all future voice access applications in the broad-band wireless environment. The various speech processing functionalities, on the other hand, can be selectively integrated in an application-dependent fashion.

A. Large-Vocabulary Continuous Speech Recognition

Large-vocabulary continuous speech recognition is the core of basic technologies for spoken language processing. The framework was developed as early as the 1970s [48]–[50], and can be represented by the simplified block diagram shown in Fig. 2. An input speech signal $x(t)$ is first represented by a sequence of feature vectors $X = (x_1, x_2, \ldots, x_t, \ldots, x_T)$, where each feature vector x_t is composed of a set of feature parameters extracted from a frame of the speech signal $x(t)$ obtained with a signal window located at time t. Today, the most commonly used feature parameters are the Mel-frequency cepstral coefficients [51] plus dynamic features [52], although very good results have also been reported using perceptual weighted linear prediction coefficients [53]. The goal of speech recognition is to find, based on the observed speech signal $x(t)$, a most likely sequence of words from a vocabulary.

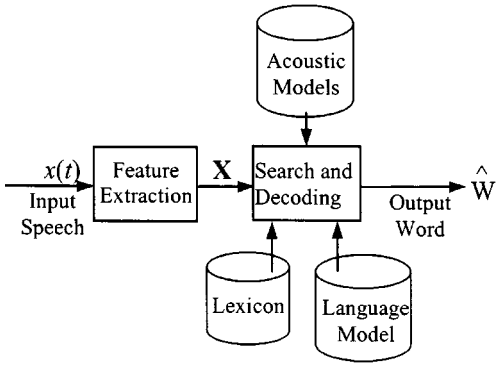

Fig. 2. Basic framework for large-vocabulary continuous speech recognition.

This goal can be achieved by finding a word sequence $\hat{W}$ that maximizes the likelihood $P(W|X)$, i.e.,

$$\hat{W} = \arg\max_{W} P(W|X) \tag{1}$$

where $W = (w_1, w_2, \ldots, w_k, \ldots, w_N)$ represents a word sequence with w_k being the kth word, and the maximization is performed over all possible word sequences. Applying Bayes' rule, we have

$$\begin{aligned}\hat{W} &= \arg\max_{W}\left[\frac{P(X|W)P(W)}{P(X)}\right] \\ &= \arg\max_{W}[P(X|W)P(W)]\end{aligned} \tag{2}$$

since $P(X)$ is the same for all word sequences W. Given the observed signal $x(t)$, the problem can be reduced to finding the word sequence $\hat{W}$ that maximizes the product of $P(X|W)$ and $P(W)$. The former is the likelihood of observing the feature vector sequence X given a word sequence W, while the latter is the *a priori* probability of observing the word sequence W, which is independent of the observed speech signal $x(t)$. This maximization process is performed using three knowledge bases as shown in Fig. 2: acoustic models, lexicon, and language models.

For the purpose of speech recognition, every word in the vocabulary is represented as a sequence of basic sounds or phones in the lexicon. Each phone is in turn represented by a statistical signal model called the hidden Markov model (HMM) [54]–[57]. As shown in Fig. 3, an HMM consists of a Markov chain with state transition probabilities a_{ij}. Each state j in an HMM is assigned an output distribution $b_j(x_t)$ that models the likelihood of observing a speech feature vector x_t at that state. The most commonly used form of $b_j(x_t)$ is the multivariate mixture Gaussian given by

$$b_j(x_t) = \sum_{m=1}^{M} C_{jm} N(x_t;\, \mu_{jm}, \mathbf{\Sigma}_{jm}) \tag{3}$$

where C_{jm} is the weight of the mixture component m in state j and $N(x;\, \mu, \Sigma)$ is a multivariate Gaussian of mean μ and covariance Σ. The state transition probabilities a_{ij} are used to model the durational variability in speech signals, while the output distributions $b_j(x_t)$ are used to model the characteristic feature variability of the speech signals. The HMMs

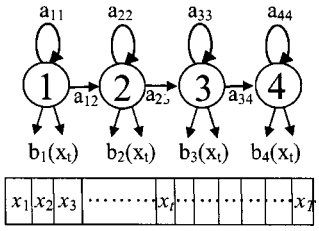

Fig. 3. Hidden Markov model.

for all the necessary phones for the desired vocabulary form the acoustic models in Fig. 2. The necessary statistical parameters for describing these HMMs are obtained from some training speech data. For a typical word sequence W, the likelihood $P(X|W)$ in (2) is evaluated from a composite model formed by concatenating the HMMs of the phones corresponding to the word sequence W.

On the other hand, the *a posteriori* probability $P(W)$ in (2) is obtained from the language model [58]–[60]. The simplest yet effective language model is the n-grams, in which it is assumed that the appearance of a word w_k depends on the preceding $n-1$ words, i.e.,

$$\begin{aligned}&P(w_k|w_1, w_2, \ldots, w_{k-1}) \\ &\quad = P(w_k|w_{k-n+1}, w_{k-n+2}, \ldots, w_{k-1}).\end{aligned} \tag{4}$$

The right-hand side of (4) is called the n-gram probabilities, which can be estimated from frequency counts in a large training text data and stored in a lookup table. However, not all n-gram probabilities can be accurately estimated from the training text data; therefore, various smoothing techniques have been developed. The probability $P(W)$ needed in (2) can then be evaluated from the n-gram probabilities. With the acoustic models providing $P(X|W)$ and the language model providing $P(W)$, the identification of the most likely output sentence $\hat{W}$ can be performed using readily available search algorithms.

B. Relevant Issues for Large-Vocabulary Continuous Speech Recognition

Although large-vocabulary continuous speech recognition has been extensively researched over the past two decades, its application to broad-band wireless presents many issues relevant to acoustic and language modeling. Solutions to these issues are very important for voice-enabled information access in the broad-band wireless environment, and are currently undergoing active research. Typical examples of these issues and possible solutions are summarized in the following.

Coarticulation is a significant challenge in acoustic modeling because it seriously affects the characteristics of naturally spoken speech signals. Context-dependent models are found to be very useful in dealing with the coarticulation effects. The simplest and most common approach is the triphone, in which every phone has a distinct HMM model for every different left and right neighbor. However, this technique leads to a large number of HMMs and a huge number of parameters to be estimated during the training process. Many approaches [61]–[65], including tied-

mixture, state-tying, phone-based component tying, and decision trees, have been developed to maximize the quantity of data available to estimate each parameter. Another major issue relevant to acoustic modeling is the mismatch between the statistical characteristics of the training speech signals and the application speech signals. The training speech signals are used for estimating the parameters of the HMMs, while the application speech signals are used for evaluating the likelihood $P(X|W)$. Mismatch between the two leads to inaccurate modeling, which, in turn, results in poor recognition performance. Two major causes for this mismatch can be identified in the context of voice-enabled information access for the broad-band wireless environment. The first, and perhaps the most challenging, cause is speaker variability. Speech signals produced by different speakers have different statistical characteristics, but the recognition system is very often trained by speech produced by people other than the user. Currently, the most important technique for dealing with speaker variability is speaker adaptation, i.e., using very limited speech data from a specific user to adapt the acoustic models obtained using training data from a larger number of speakers. Substantial work has been done in this area, and many different approaches have been developed. Typical examples include the global transformation for all acoustic model parameters, clustering the Gaussian distributions and defining individual transformation for each cluster, locally smoothing the updated parameters based on available adaptation data, and the classical MAP estimation for the parameters if the available adaptation data are adequate [66]–[72]. Considering the large number of users in the future global wireless network environment and the difficulties of collecting speech data for each user, it is foreseeable that speaker variability will remain a very important issue in large-vocabulary continuous speech recognition. The second cause, referred to as environmental variability, is the mismatch between the environments in which the training and the application signals are obtained. With the mobility provided by broad-band wireless communications, it is impossible to guarantee that the environment in which voice-enabled information access is used will be the same as the environment in which the acoustic models are trained. Therefore, speech recognition technologies for broad-band wireless applications must deal with the mismatch due to variations in microphone reception (hands free or handheld), room acoustics (in a car or in a building), mode of operation (driving or walking), broad-band wireless transport technology (outdoor cellular system or indoor wireless LAN), background noise (on a busy street or in a quiet room), and applications. Distortions due to the wireless communication channel, microphone, and room acoustics can be approximately modeled as linear distortion, which is represented by the convolution of the speech signal with some unknown function. The background noise, on the other hand, can be represented as an additive noise that further corrupts the signal. All these conditions vary for different applications. Since it is impossible to collect training data for all different cases, many approaches have been developed to improve recognition robustness against environmental variability. For example, since convolution in the time domain introduces an additive component in the cepstral feature parameters, linear distortion can be effectively removed if the additive component can be estimated with reasonable accuracy [73]–[75]. On the other hand, noise-robust features can be extracted by various techniques. The acoustic models can also be modified to include the environmental changes [76], [77]. Considering the completely uncontrollable acoustic conditions in the future global wireless network environment, robustness against environmental variability will be very critical. Furthermore, most current works on speech recognition are based on read or prepared speech data, i.e., the speaker reads or speaks from some prepared text when the speech signals are collected. However, in future broad-band wireless applications, the speech signals to be handled will be spontaneous, i.e., speech produced spontaneously without a prepared text. There exist significant differences between the two, primarily due to the increased coarticulation and fluency, variable speaking rate, and various types of disfluency such as hesitations, false starts, and repairs in spontaneous speech. Experimental results indicate that recognition accuracy can be significantly degraded when spontaneous speech is to be recognized. Active research is in progress in this area, and encouraging results are being produced from time to time.

In language modeling, on the other hand, the capabilities of the n-grams are generally too limited to be useful in the broad-band wireless environment because they only describe the local behavior of a language. Better models, e.g., tree-based models, trellis models, trigger models, history models [78]–[82], and word-class-based models, are available. A new approach to include more semantic information based on a word–document matrix has also been found to be very successful [72], [83]. As mentioned earlier, the parameters of the n-grams or the more advanced language models are obtained from training texts. Just as in acoustic modeling, the most difficult challenge is the mismatch in statistical characteristics between the training text and the application text, which comes from the high variability in human languages. For example, the n-gram probabilities in (4) estimated from some training text on one subject domain may not be very helpful in speech recognition applications in another. Users of future voice applications for the broad-band wireless environment are expected to access information from a plethora of diverse subject domains at any time, therefore, the ability to adapt the language models to the right subject domains based on very limited input data is very important. A good example of a successful adaptation approach [84] is the cache-based approach. In addition to statistical mismatch, texts from different subject domains usually contain different vocabularies. In the basic framework in Fig. 2, recognition is primarily based on a lexicon. The lexicon can never include all possible words, and even if it did, the search space would be too large for practical implementations. If the input speech signal includes some words not in the lexicon, the recognition process will produce some errors. A promising approach for handling this problem is to generate dynamic lexicons by

automatically extracting contents from networked resources and classifying them into different subject domains [85]. Finally, because of the size of the search space for the input speech given the lexicon, acoustic models, and language models, efficient search techniques have always been desirable. This is especially true when speech recognition is applied in a broad-band wireless communication system because of the mobility and increased diversification in application scenarios. Typical important search techniques include depth-first search and breadth-first search, stack decoding, A^*-decoding, Viterbi decoding, beam search, and look-ahead approaches [72], [86].

C. Speech Understanding

The speech recognition technologies mentioned above simply transform the speech signal $x(t)$ into a word sequence $\hat{W}$ without understanding the meanings carried by the word sequence. Speech understanding means the extraction of the meaning of the speech signal $x(t)$ so that the machine or network can understand the intention of the speaker and perform appropriate functions accordingly. Speech understanding is a key element for broad-band wireless applications, especially in the voice access to personalized intelligent agents and voice retrieval of information mentioned above. In general, understanding the meaning of an arbitrary sequence of clearly identified words by machines is a very difficult problem because the general knowledge needed for understanding human language is almost unlimited. Understanding speech signals is even more difficult because of the added ambiguities in the word sequence caused by the inevitable recognition errors. However, language or speech understanding for a specific task domain is achievable because only the knowledge relevant to the particular task domain is necessary. For example, in a train schedule information task, the word sequence "from London to Paris" is not too difficult to understand, as long as the system knows *a priori* that "London" and "Paris" are cities, the city after the word "from" is the origin of the trip, and the city after the word "to" is the destination. In this case, the origin and destination can be defined as two "slots," and understanding is completed when these two slots are filled in with city names. In general, there can be a wide variety of scopes and degrees of difficulties for speech understanding depending on the purposes, applications, and task goals. Similarly, there are also a wide variety of approaches and strategies developed for speech understanding with different purposes, applications, and task goals. The appropriate choice of speech understanding technologies for broad-band wireless applications depends on the application scenario.

Language understanding technologies for printed texts and sentences are typically grammatical. A parsing process is used for generating parsing trees for the sentences in order to analyze the syntactic and semantic relationships among the words and phrases so that the meaning of the sentences can be interpreted accordingly. Very often, the necessary knowledge to be considered includes not only the syntactic and semantic parts, but also the pragmatic and discourse knowledge, history of the interaction, and knowledge regarding the task domain [87]. Speech understanding for recognized word sequences, on the other hand, cannot be achieved this way because of the recognition errors and ambiguities as well as the nongrammatical forms in the recognized word sequences. As a result, complete parsing of the word sequences is very often unachievable. However, in a limited task domain with adequate domain knowledge, very often the partial results of parsing small portions of the recognized word sequences or phrases, or the detection of some key phrases or keywords, may lead to reasonable understanding of the speech signals. It is also possible to define the "language" and develop or train deterministic or stochastic grammars for some specific task domains based on the domain knowledge [88], [89].

On the other hand, understanding is also achievable in some cases via stochastic models similar to the speech recognition framework mentioned previously. For example, a set of "concepts" may be defined for the task domain, and a concept structure $C = [c_1, c_2, \cdots c_k, \cdots, c_L]$ may be used to specify the meaning of a word sequence where c_k is a concept or another structure of concepts. In this case, the goal of understanding is to find the most likely concept structure $\hat{C}$ given the recognized word sequence W. Mathematically, we have

$$\hat{C} = \arg\max_{C} P(C|W) \tag{5}$$

where $P(C|W)$ is the *a posteriori* likelihood of C given the recognized word sequence W. Equations (5) and (1) are of the same form, thus (5) can be handled similarly to (1) to yield

$$\begin{aligned}\hat{C} &= \arg\max_{C}\left[\frac{P(W|C)P(C)}{P(W)}\right] \\ &= \arg\max_{C}[P(W|C)P(C)].\end{aligned} \tag{6}$$

Here, $P(C)$ can be estimated by a "concept language model" that can be n-grams of concepts trained from some corpora, and $P(W|C)$ can be estimated from a "lexical realization model" that specifies how words are generated from the meaning or concept structures. Many approaches have been developed along this direction, and some of them can be integrated with the grammatical approach mentioned above [90]–[92].

There are many other alternative approaches and strategies, among them WordNet is perhaps one worth mentioning. In this approach, lexicalized concepts are represented by sets of synonyms, or the "synsets." Semantic relationships are then represented by links among the "synsets," and these links form a "WordNet." A word may belong to more than one "synset" if it has more than one meaning. WordNet is a very powerful knowledge base for both language and speech understanding [90], [93].

D. Speaker Verification

User authentication is crucial when accessing a private information database or a personalized intelligent agent. As mentioned previously, the traditional mode of user authentication is impractical for miniature wireless terminals.

Therefore, it can be envisioned that user authentication using speaker verification will be a key element for voice access of global information in the broad-band wireless environment. In speaker verification, the user claims an identity using voice, and the system or the network decides whether the claim should be accepted or rejected based on the speech signals received. Speaker verification is therefore a statistical pattern recognition problem. In a speaker verification system, feature parameters are extracted from the input speech to form a sequence of feature vectors just as in large-vocabulary speech recognition. However, here the purpose is to verify the speaker rather than to determine the word sequence; therefore, the acoustic models, lexicon, and language models in Fig. 2 are very often not necessary. Instead, a database for the statistical distributions of the feature vectors in the feature space for speakers to be considered is used here. The accuracy of speaker verification is measured by the rates of false acceptance and false rejection. The tradeoff between these two rates can usually be adjusted by tuning one or more parameters in the verification algorithm. The requirements for these two rates can be quite different for different applications. For example, when speaker verification is used to control access to credit card accounts, false acceptance of an intruder may result in a disaster, but false rejection of a legitimate user is not as detrimental because the user can always try again and get through the next time. However, if speaker verification is used to control access to medical databases, false rejection of a legitimate user may cause delays in emergency medical treatment, which can have serious consequences.

There are several different types of speaker verification techniques [94]–[96]. The simplest case uses a fixed password, i.e., the user is required to speak his/her password when claiming an identity. Assuming that one or more utterances from the user for this password have been used in the training process, it is not too difficult to verify the speaker. Therefore, this type of speaker verification technique generally achieves the highest accuracy (lowest false acceptance and false rejection rates). However, it is also the easiest to break. For example, an intruder can impersonate a legitimate user by playing a recording of the password. Speaker verification systems for the broad-band wireless environment are especially vulnerable to this type of attack if encryption is not properly designed or implemented in the wireless link, because a potential intruder may eavesdrop on the wireless link to obtain the necessary recordings. A more secure type of speaker verification approach is to ask the speaker to speak a randomly assigned word sequence. This is much more difficult to break, but the technologies involved are also much more complicated. For instance, the trajectories in the feature space through which feature vectors move with time are completely different for different word sequences. A sophisticated verification algorithm is therefore necessary to verify against all the possible trajectories. In general, it is desirable to select feature parameters with the highest discriminating power for speaker verification. Furthermore, when applied to broad-band wireless communications, these feature parameters as well as the verification algorithm must also be robust against channel impairments and environment variability, so that the same verification accuracy is maintained in different application scenarios.

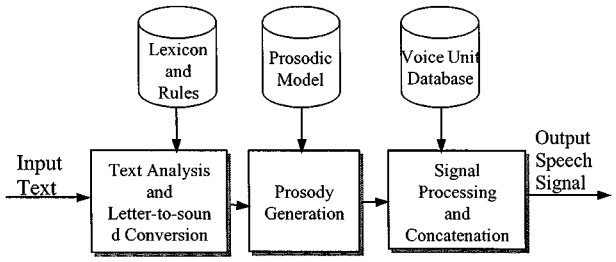

Fig. 4. Key elements of the TTS synthesis technologies.

E. Text-to-Speech (TTS) Synthesis

TTS synthesis is a technology for converting an arbitrary text into speech signals. This is the core technology for providing textual information to users in the form of voice. For example, people can listen to e-mails and web pages read by TTS systems over mobile handsets. The key elements for typical TTS synthesis technologies are shown in Fig. 4. The input texts are first analyzed with the help of a lexicon plus such knowledge as part-of-speech information and other linguistic rules to obtain the general structure of the sentences. Letter-to-sound rules are also used to generate a sequence of the phone units to be produced by the system. The prosody (such features as fundamental frequency, intonation, duration, and energy for each phone unit and so on) for the sentence is then generated from a prosodic model. The voice units necessary for the desired sentence are then selected from a voice unit database. If the selected voice units have different prosodic features from those desired, some modifications on the voice units have to be made using signal processing techniques. The voice units are finally concatenated and smoothed to produce the output speech signal. The most difficult part here is generating the right prosody such that the output speech sounds natural, which relies on the availability of a very good prosodic model. The performance of TTS synthesis technologies is primarily characterized by two factors: the intelligibility and the naturalness of the output speech signal. The former is the key for the technologies to be useful, and has been achieved very well in general. The latter depends very much on the prosodic model, and is much more difficult because the prosody of human speech is difficult to model. Today, some TTS technologies developed at several organizations provide synthetic speech with very good naturalness, though many others are not necessarily satisfactory in this respect [97], [98].

In the past, most TTS systems produced the desired output speech by concatenating selected sets of voice units that were pre-stored in a database. Recently, a new paradigm of "corpus-based" TTS synthesis with online unit selection has emerged. This provides much better output speech quality at the cost of much more storage and computation requirements. In this approach, a large speech corpus (on the order of 10 h or even much more of speech) produced by a single speaker is collected. The corpus is designed so that almost

all linguistic and prosodic features for the target language (either for general domain or specific domain) have been included. Parallel analysis of all prosodic and linguistic features of the speech signals as well as the corresponding texts can lead to a much better prosodic model. There can be many repetitions of a given voice unit in the corpus, but in different context with different prosodic features. During the synthesis process, the most appropriate units, longer or shorter, with the desired prosodic features within the corpus are automatically retrieved and selected online in real time, and concatenated (with modifications when necessary) to produce the output speech. In this way, very often longer units (especially commonly used words or even phrases) can be used in the synthesis if they appear in the corpus with desired prosodic features. Also, in this way the need for signal modification to obtain desired prosodic features for a voice unit, which usually degrades the naturalness of speech, is significantly reduced. This is why much better performance can be achieved using this approach.

V. Speech Processing Functionalities for the Broadband Wireless Environment

The basic technologies mentioned in the previous section can be integrated to implement many speech processing functionalities that are necessary for voice access of information in the broad-band wireless environment. Typical examples of these functionalities include dictation and transcription, audio indexing and retrieval, spoken dialogue, and multilingual functionality.

A. Dictation and Transcription

Dictation refers to the production of written documents via voice input. It is the most obvious application of the large-vocabulary continuous speech recognition technologies, and is the key functionality of remote authoring mentioned above. In dictation, the voice input is usually assumed to be prepared speech, which is very close to read speech. Furthermore, a known speaker who may have tried to adapt the acoustic and language models usually produces the input speech. Many good systems for a variety of languages exist today, and a good number of products for personal computers are commercially available. In particular, some special products for professional applications, such as medical use or legal use, have been quite successful. For almost all commercial products, the microphone used by the users is selected and provided by the system developer. All these indicate that constraints such as known speaker, close to read speech, adapted acoustic and language models, assigned subject domain, and specified microphone characteristics, are imposed to control the application environment in order to achieve better recognition results. These constraints are reasonable and acceptable for today's applications. However, they must be relaxed in order to meet the requirements of broad-band wireless applications. Unfortunately, there still exists a gap between the performance of current technologies and the users' expectation [72].

Transcription is similar to dictation except that the speaker does not necessarily intend to produce the written documents. In transcription, the voice is produced in more natural scenarios such as telephone conversation, broadcast news, meeting recordings, etc., while the text version of the voice is needed for some other applications such as topic spotting and audio indexing. Transcription is useful in such applications as retrieval of voice messages, voice files, or notes from a voice notebook, producing meeting and interview minutes, etc. In general, the speech for such applications is obtained under much less controlled conditions and accurate transcription is a much more challenging task. Usually, before large-vocabulary continuous speech recognition can proceed, it is necessary to perform automatic segmentation of the speech signals, i.e., the partition of the speech signals into signal segments with acoustically homogeneous conditions such as produced by the same speaker, under the same background noise conditions, collected with the same bandwidth and so on, so that different adaptation techniques can be applied to different segments and better transcription results can be obtained. Currently, the word error rates for the transcription of broadcast news on the order of $\sim$15%–25% for several languages, including American English, French, German, Spanish, and Chinese, have been achieved. This particular word error rate is already quite useful for applications such as audio file indexing and retrieval. However, the achievable word error rates for telephone conversation are significantly higher, apparently due to the much more spontaneous nature of the speech signals and the uncontrolled, much wider subject domains [72]. It can, therefore, be seen that implementing the transcription functionality for voice-enabled information access in the broad-band wireless environment is even more challenging, due to the increased variations in application scenarios.

B. Audio Indexing and Retrieval

As information accumulates in the global network, information indexing and retrieval become important technologies. Currently, most of the work in this area has been focused on the indexing and retrieval of stored text. Related work on stored video has also become active in recent years. As the technologies for spoken language processing rapidly progress, indexing and retrieval for audio signals, in particular speech signals such as news broadcasts, meeting recordings, and voice messages, become crucial for many emerging applications. The purpose of audio indexing and retrieval is to create a concise structural summarization of the stored audio signals in terms of, e.g., stories, topics, speakers, etc., and to construct an efficient database such that the audio signals can be easily retrieved using queries on, e.g., events, people, organizations, locations, etc. The application in future broad-band wireless environment as mentioned above is obvious.

The technologies of large-vocabulary continuous speech recognition and the transcription functionality mentioned in the previous subsection are the key elements for audio indexing and retrieval, but some extra elements are also necessary. First, name-spotting technologies that extract

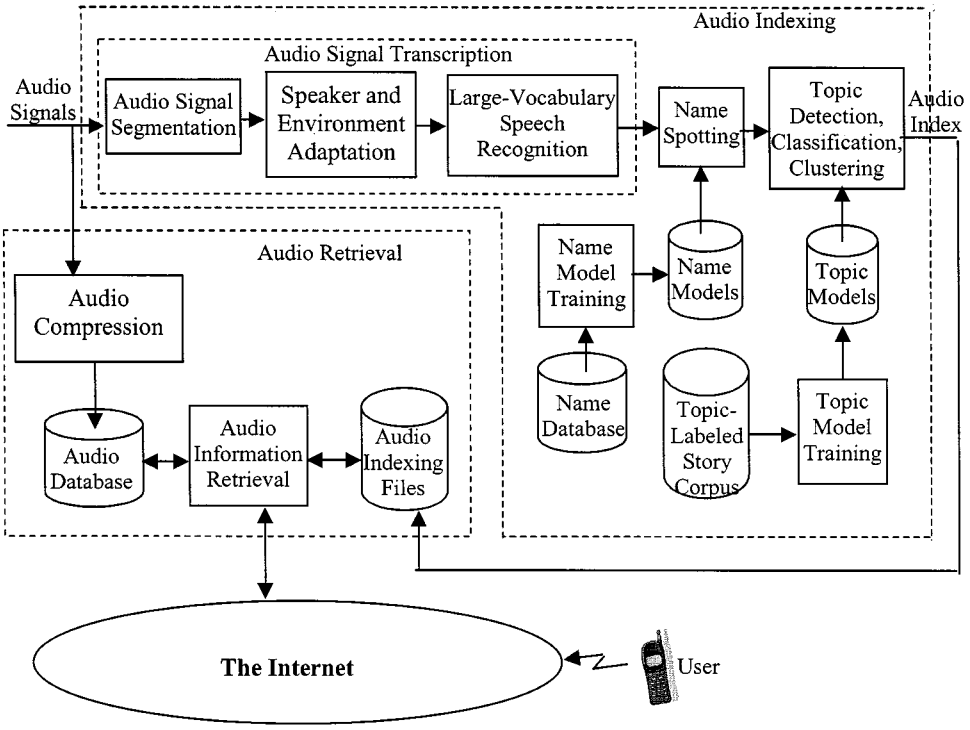

Fig. 5. Example simplified block diagram for processes of audio indexing and retrieval.

names of persons, organizations, locations, etc., are needed because these names are usually the keywords or phrases for the retrieval of information from the voice files. Some statistics-based plus lexicon-based approaches have been useful when applied on the transcribed texts. For example, HMMs for the named entities based on the words in the texts have been found to be very successful. Second, topic detection and classification technologies are needed. These technologies identify the topics of the stories based on statistical analysis of the words in the texts, and then cluster the stories based on their topics. This is usually achieved by a set of training documents with manually identified topics, and in most cases a story can have several topics. Very often some probabilistic models are used. An example simplified block diagram for such audio indexing and retrieval processes is shown in Fig. 5, where the audio indexing including transcription, name spotting, and topic processing are at the top of the figure, while audio retrieval accessed by the user via the network is shown at the bottom. All of these are similar to the retrieval of text information, but extra difficulties and challenges do exist in the retrieval of audio information due to the fundamental differences in the nature of the text and audio information. For example, the audio news broadcast needs to be automatically segmented, and the segmentation errors may cause some problems; but for text information all documents are separated with clear boundaries and segmentation errors are unlikely. Also, the inevitable errors in the transcribed texts lead to additional problems that are not present in the retrieval of text information. For example, in texts, proper nouns are usually easily identified by the capital letters, but for audio information they may not be included in the lexicon of the large-vocabulary speech recognition system; thus, some errors may be produced instead. Technologies to develop dynamic lexicons and language models updated with the dynamic information to be handled are therefore highly desirable in the future [99].

C. Spoken Dialogue

Spoken dialogue systems, or voice conversational interfaces, refer to systems or interfaces that allow the users to interact with a machine or network using voice to retrieve information, conduct transactions, or perform other tasks. The application in the broad-band wireless environment is obvious. For instance, the user may invoke the interactive functions of the personalized intelligent agents using the wireless handset to instruct an agent to take some actions on the user's behalf. Because the user's intention may not be clearly specified in a few commands or sentences, the most convenient and user-friendly approach is to allow the user to dialogue with the system or network. In the past decade, many such systems have been successfully developed for a wide variety of applications using different design principles and architectures. In the simplest case, the user may simply utter a few isolated words such as "YES" or "NO" during the dialogue, while in the most complicated case the user may use unconstrained input voice to direct the system to handle some desired tasks within a reasonable domain. The "initiative" is usually an important parameter for indicating the flexibility of the conversation the system is capable of handling. One extreme in this dimension is the "system-initiative" dialogue system, in which the user is asked a sequence of questions regarding the information needed to perform the task, and the user needs to respond to each question with a valid

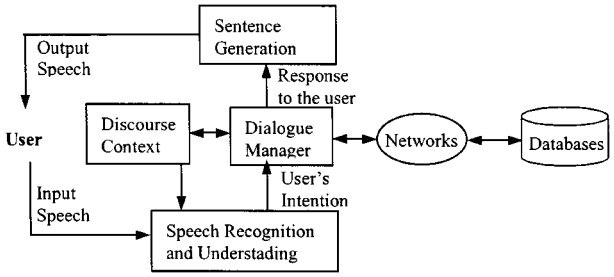

Fig. 6. General architecture of a spoken dialogue system.

voice input. Such systems are already available for many applications, although they may not be efficient enough in the broad-band wireless environment. The other extreme in this dimension is the "user-initiative" dialogue system, in which the user is free to take the initiative during the dialogue and speak with the system in whichever way he prefers. In the task of travel arrangements, for example, the user may reserve airline tickets and hotel rooms as if talking to a human travel agent. The technologies for implementing a user-initiative dialogue system do not yet exist today. Achievable today are the "mixed-initiative" dialogue systems which are in between the above two extremes, i.e., the system uses some prompts to direct the dialogue and indicate the information needed for the task, while the user can also initiate some reasonable questions or instructions to express his intention. Of course, systems that are closer to the "system initiative" extreme are more reliable; while systems that are farther from that extreme are more difficult to implement [100]–[104].

Although the design principles of a spoken dialogue system may vary significantly from case to case, in most situations, the simplified block diagram shown in Fig. 6 can represent its general structure. Speech recognition and understanding (see Section IV) are first performed on the input speech. Speech understanding for dialogue systems can be performed independently of or jointly with speech recognition. The dialogue manager then accepts the user's intention extracted from the input speech and performs the desired tasks, and finally produces some responses to the user. During a dialogue, some concepts relevant to the subject of the conversation may remain valid or unchanged throughout the dialogue, but are mentioned only in one of the sentences. With the aid of the discourse context, the system can understand that these concepts remain valid even if they are not mentioned in the following sentences. There are many possible strategies for designing the dialogue manager. For example, sub-dialogues can be used for the confirmation of the understood concepts, error recovery, reduction or expansion of the scope of the user's request, clarification of the ambiguities, etc. [105]–[107]. The response to the user is finally formulated as sentences and produced as speech signals to be transmitted to the user. The dialogue system performance is usually evaluated by such parameters as task success rates, average dialogue turns, and dialogue efficiency. However, developing a generalized framework for evaluating and analyzing the performance of a spoken dialogue system is still an open problem. The portability of the spoken dialogue technologies across many different tasks in different domains is another open problem in the study of dialogue systems. It is believed that a set of technologies including a universal platform plus a set of convenient tools with high portability is desired, but still does not exist today [108].

D. Multilingual Functionality

A special feature of the future world of global information networks is that the information activities and knowledge systems of all parts of the world will be completely integrated by the networks, and the information handled by the networks will be completely globalized, multicultural, and multilingual. As a result, when a user tries to access the network information via voice in the broad-band wireless environment, not only the language he uses in his voice can be one of the many languages in the world, but the information contents over the networks are represented in the many different languages in the world as well. The total number of languages considered really depends on the population of the users and the languages used in the contents over the networks. For example, there are at least dozens of "important" languages used by a large enough population, and at least 20 of them have been studied for computer processing of language or speech. But some people believe there are at least 4000 different languages all over the world. The number of different languages also depends on how dialects and languages are differentiated. For example, the Chinese language is spoken by roughly a quarter of the world's population. Very often it is considered as a single language, but actually many of its dialects sound completely different, and even have different lexicons. Some people believe that the many dialects of Chinese language can be considered as hundreds of different languages, each of which is actually spoken by a large enough population. As an example, when a user tries to retrieve some information via voice, the most relevant information over the networks may be in many languages other than the language he is using. This is the very important problem of "cross-language information retrieval" in the area of information retrieval. A typical scenario is shown in Fig. 7. The language used for each piece of information obtained from the multilingual resources in the Internet is identified, and then processed in each respective language. Machine translation is often needed for interfacing with the user. The problem of different languages caused by global roaming in wireless environments mentioned previously is another example. All these considerations lead to the need for multilingual functionality.

In earlier years, some people believed the general principles of the spoken language processing technologies were language independent; therefore, all that is needed is to collect enough linguistic data so as to construct acoustic and language models and lexicons for each target language. This is called the "localization" of technologies to different languages—an issue considered to be trivial and of no scientific value. After carefully analyzing many different languages, it is now realized that although many of the general principles of the technologies are indeed language independent, all the principles are definitely not equally extensible and efficient

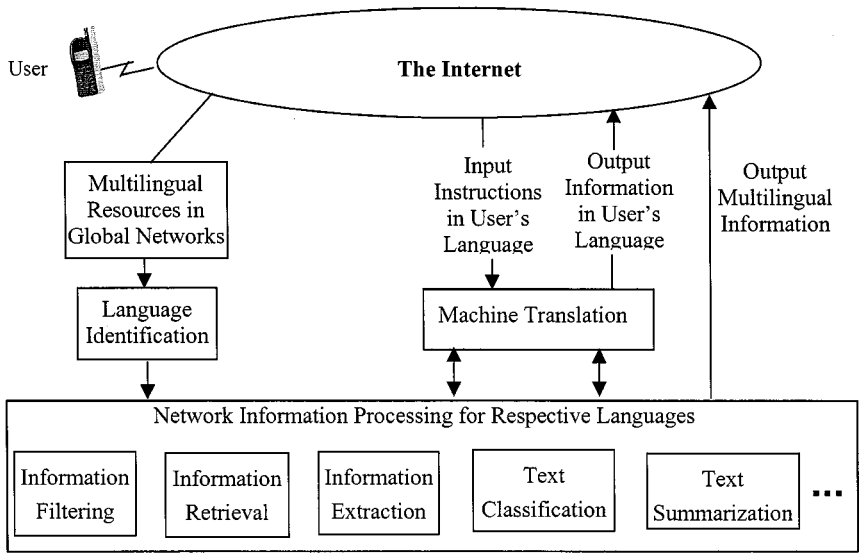

Fig. 7. Typical scenario for cross-language network information processing.

電腦科技的進步改變了人類的工作與生活方式

(Advances in computer technologies have changed the working and living styles of human beings.)

Fig. 8. Example Chinese sentence showing the fact that the "words" in Chinese are not well defined. Each line segment here represents an acceptable word composed of one to several characters.

to all different languages due to the special characteristics of the different languages. In addition, very often the different languages do bring new challenges not encountered before in the "mainstream" languages such as English [109]. For example, in the Chinese language, each character has its own meaning and can play some linguistic role independently. Although a Chinese word is composed of one to several characters, in a written sentence there are no blanks between the words. As a result, the segmentation of a Chinese sentence into words is not unique, there does not exist a commonly accepted "lexicon," and in fact the "word" in Chinese is not very well defined. An example is shown in Fig. 8, in which the string of Chinese characters is a sentence, and each line segment represents an acceptable "word." This is one reason why extending the previously mentioned "word-based" technologies developed based on western alphabetic languages (such as the large-vocabulary continuous speech recognition shown in Fig. 2 based on a lexicon of words and the language model defined by relations among words) to the Chinese language may not be trivial. In another example, both Thai and Chinese languages are tonal, i.e., the tones lead to lexical meaning. As a result, the recognition of the tones becomes an important issue though this does not exist at all in English [110], [111]. Similar situations have been found in quite a few other languages [109]. Two typical approaches currently under active research for providing the multilingual functionality are worth mentioning here. In the first approach, the construction of a set of "language-universal" acoustic models is considered, such that this single set of acoustic models can be used in the recognition of as many languages as possible. The second approach is the automatic translation among the speech signals of major languages. Such speech translation technologies may include the integration of large-vocabulary speech recognition, machine translation and TTS synthesis technologies, or complicated integrated models [72], [109], [112]. Both of these approaches will be very important for providing the multilingual functionality necessary to support global roaming in the broad-band wireless environment.

VI. Conclusion

A vision of the integration of broad-band wireless and spoken language processing technologies is presented in the context of voice access to the global information network. Emerging broad-band wireless technologies are surveyed, and potential applications of spoken language processing technologies in the broad-band wireless environment are identified. The technical backgrounds, including the fundamentals of spoken language processing technologies as well as technical challenges that arise due to the unique nature of broad-band wireless communications and possible solutions, are presented in detail. It is believed that a good integration of spoken language processing and broad-band wireless technologies is an extremely attractive way to achieve the goal of "anytime, anywhere" access to the information infrastructure. Although many difficult technical hurdles and a substantial amount of work lies ahead, we believe that the practical demands from the users, the technical interests of the technology developers, and the business motivation of the service providers will soon merge to realize the vision.

References

[1] A. Furuskar *et al.*, "EDGE, enhanced data rates for GSM and TDMA/136 evolution," *IEEE Pers. Commun.*, vol. 6, pp. 56–67, June 1999.

[2] R. van Nobelen *et al.*, "An adaptive radio link protocol with enhanced data rates for GSM evolution," *IEEE Pers. Commun.*, vol. 6, pp. 54–64, Feb. 1999.

[3] N. Sollenberger *et al.*, "The evolution of IS-136 TDMA for third-generation wireless services," *IEEE Pers. Commun.*, vol. 6, pp. 8–19, June 1999.
[4] R. Pirhonen *et al.*, "TDMA based packet data system standard and deployment," in *1999 IEEE Vehicular Technology Conf.*, pp. 743–747.
[5] M. Austin *et al.*, "Service and system enhancements for TDMA digital cellular systems," *IEEE Pers. Commun.*, vol. 6, pp. 20–33, June 1999.
[6] D. Terasawa and E. G. Tiedemann Jr., "CdmaOne (IS-95) technology overview and evolution," in *Proc. 1999 IEEE Radio Frequency Integrated Circuits Symp.*, pp. 213–216.
[7] D. Knisely *et al.*, "Evolution of wireless data services: IS-95 to CDMA2000," *IEEE Commun. Mag.*, vol. 36, pp. 140–149, Oct. 1998.
[8] T. Ojanpera *et al.*, "An overview of third-generation wireless personal communications: A European perspective," *IEEE Pers. Commun.*, vol. 5, pp. 59–65, Dec. 1998.
[9] "Special issue on IMT-2000 standards efforts of the ITU," *IEEE Pers. Commun.*, vol. 4, Aug. 1997.
[10] J. Uddenfeldt, "Digital cellular—Its roots and its future," *Proc. IEEE*, vol. 86, pp. 1319–1324, July 1998.
[11] V. Garg *et al.*, "Third generation (3G) mobile communication systems," in *Proc. 1999 IEEE Conf. Personal Wireless Communications*, pp. 39–43.
[12] E. Dahlman *et al.*, "UMTS/IMT-2000 based on wideband CDMA," *IEEE Commun. Mag.*, vol. 36, pp. 70–80, Sept. 1998.
[13] A. Sasaki *et al.*, "The current situation of IMT-2000 standardization activities in Japan," *IEEE Commun. Mag.*, vol. 36, pp. 145–153, Sept. 1998.
[14] T. Rahman *et al.*, "The cellular phone industry in Malaysia: Toward IMT-2000," *IEEE Commun. Mag.*, vol. 36, pp. 154–156, Sept. 1998.
[15] C. Shumin, "Current development of IMT-2000 in China," *IEEE Commun. Mag.*, vol. 36, pp. 157–159, Sept. 1998.
[16] K. Wee and Y.-S. Shin, "Current IMT-2000 R&D status and views in Korea," *IEEE Commun. Mag.*, vol. 36, pp. 160–164, Sept. 1998.
[17] "TR45 proposed RTT submission (UWC-136)," TR-45.3/98.03.03.19, Mar. 1998.
[18] "The Cdma2000 ITU-R RTT candidate submission," TR-45.5, May 1998.
[19] "International telecommunication union press release," ITU/99-22, Nov. 5, 1999.
[20] J. Mitola, "The software radio architecture," *IEEE Commun. Mag.*, vol. 33, pp. 26–38, May 1995.
[21] J. Razavilar *et al.*, "Software radio architecture with smart antennas: A tutorial on algorithms and complexity," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 662–676, Apr. 1999.
[22] Y. K. Yoon and M. Ulema, "A wireless local loop system based on wideband CDMA technology," *IEEE Commun. Mag.*, vol. 37, pp. 128–135, Oct. 1999.
[23] M. Progler *et al.*, "Air interface access schemes for broadband mobile systems," *IEEE Commun. Mag.*, vol. 37, pp. 106–115, Sept. 1999.
[24] K. Chua *et al.*, "Wireless broadband communications: Some research activities in Singapore," *IEEE Commun. Mag.*, vol. 37, pp. 84–90, Nov. 1999.
[25] J. Mikkonen *et al.*, "Emerging wireless broadband networks," *IEEE Commun. Mag.*, vol. 36, pp. 112–117, Feb. 1998.
[26] L. M. Correia and R. Prasad, "An overview of wireless broadband communications," *IEEE Commun. Mag.*, vol. 35, pp. 28–33, Jan. 1997.
[27] N. R. Prasad and H. Teunissen, "A state-of-the-art of HIPERLAN/2," in *Proc. 1999 IEEE Vehicular Technology Conf.*, pp. 2661–2666.
[28] J. Haine, "HIPERACCESS: An access system for the information age," *Electron. Commun. Eng. J.*, pp. 229–235, Oct. 1998.
[29] W. Honcharenko *et al.*, "Broadband wireless access," *IEEE Commun. Mag.*, vol. 35, pp. 20–26, Jan. 1997.
[30] R. B. Marks, "The IEEE 802.16 working group on broadband wireless," *IEEE Network Mag.*, vol. 13, pp. 4–5, Mar./Apr. 1999.
[31] K. Pahlavan *et al.*, "Wideband local access: Wireless LAN and wireless ATM," *IEEE Commun. Mag.*, vol. 35, pp. 34–40, Nov. 1997.
[32] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, Jan. 1996. IEEE 802.11D3.
[33] *Radio Equipment and System; High Performance Radio Local Area Network (HIPERLAN); Type 1; Functional Specification*, Dec. 1996. ETSI TC-RES.
[34] *Radio Equipment and System; High Performance Radio Local Area Network (HIPERLAN); System Definition*, July 1994. ETSI TC-RES.
[35] R. Nakatsu, "ANSER: An application of speech technology to the Japanese banking industry," *Computer*, vol. 23, pp. 43–48, Aug. 1990.
[36] World Wide Web Consortium. Voice browser activity statement. [Online]. Available: http://www.w3.org/Voice/Activity.html
[37] ——, Aural cascading style sheets (ACSS). NOTE-ACSS-970107. [Online]. Available: http://www.w3.org/Style/CSS/Speech/NOTE-ACSS
[38] Motorola Inc., "Motorola's VoxML™ voice markup language,".
[39] C. Asakawa and T. Itoh, "User interface of a home page reader," in *1998 ACM Conf. Assistive Technologies*, 1998-4.
[40] J. R. Glass, "Real-time telephone-based speech recognition in the Jupiter domain," in *Proc. 1999 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 61–64.
[41] D. Stallard, "BBN position paper on conversational web access," presented at the W3C Workshop Voice Browsers, Oct. 1998.
[42] C. T. Hephill *et al.*, "Speech-aware multimedia," *IEEE Multimedia*, vol. 3, pp. 74–78, Spring 1996.
[43] S. Goose *et al.*, "1-800-Hypertext: Browsing hypertext with a telephone," in *Proc. 1998 ACM Int. Conf. Hypertext*, pp. 287–288.
[44] M. Brown *et al.*, "PhoneBrowser: A web-content-programmable speech processing platform," in W3C Workshop Voice Browsers, Oct. 1998.
[45] M. Wynblat and S. Goose, "Toward imporving audio web browsing," presented at the , Oct. 1998.
[46] Y. Muthusamy *et al.*, "Speech-enabled information retrieval in the automobile environment," in *1999 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 2259–2262.
[47] R. V. Cox, C. A. Kamm, L. R. Rabiner, J. Schroeter, and J. G. Wilpon, "Speech and language processing for next-millennium communications services," *Proc. IEEE*, vol. 88, pp. 1314–1337, Aug. 2000.
[48] J. K. Baker, "The dragon system—An overview," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, no. 1, pp. 24–29, 1975.
[49] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE*, vol. 64, pp. 532–556, 1976.
[50] S. Young, "A review of large-vocabulary continuous-speech recognition," *IEEE Signal Proc. Mag.*, vol. 13, pp. 45–57, Sept. 1996.
[51] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 4, pp. 357–366, 1980.
[52] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 52–59, 1986.
[53] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, 1990.
[54] L. R. Bahl, P. F. Brown, P. V. De Souza, and R. K. Mercer, "Speech recognition with continuous parameters hidden Markov models," *Comput. Speech Lang.*, vol. 2, no. 3/4, pp. 219–234, 1987.
[55] B.-H. Juang, "Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains," *AT&T Tech. J.*, vol. 64, no. 6, pp. 1235–1249, 1985.
[56] L. A. Liporace, "Maximum-likelihood estimation for multivariate stochastic observations of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-28, no. 5, pp. 729–734, 1982.
[57] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[58] S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 35, no. 3, pp. 400–407, 1987.
[59] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependencies in stochastic language modeling," *Comput. Speech Lang.*, vol. 8, no. 1, pp. 1–38, 1994.
[60] G. Riccardi, R. Pieraccini, and E. Bocchieri, "Stochastic automata for language modeling," *Comput. Speech Lang.*, pp. 265–293, 1996.
[61] X. D. Huang, H. W. Hon, M. Y. Hwang, and K. F. Lee, "A comparative study of discrete, semi-continuous and continuous hidden Markov models," *Comput. Speech Lang.*, vol. 7, no. 4, pp. 359–368, 1993.

[62] M. Y. Hwang and X. D. Huang, "Shared distribution hidden Markov models for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 414–420, Oct. 1993.
[63] V. Digalakis, P. Monaco, and H. Murveit, "Genones: Generalized mixture tying in continuous speech HMM-based speech recognizers," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 281–289, July 1996.
[64] S. J. Young and P. C. Woodland, "State clustering in HMM-based continuous speech recognition," *Comput. Speech Lang.*, vol. 8, no. 4, pp. 369–384, 1994.
[65] A. Kannan, M. Ostendorf, and J. R. Rohlicek, "Maximum likelihood clustering of Gaussians for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 453–455, July 1994.
[66] M. Tonomura *et al.*, "Speaker adaptation based on transfer field smoothing using maximum a posteriori probability estimation," in *Proc. ICASSP 1995*, pp. 668–691.
[67] V. Digalakis, D. Rtischev, and L. Neumeyer, "Fast speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 357–366, Sept. 1995.
[68] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comp. Speech Lang.*, vol. 9, no. 2, pp. 171–185, 1995.
[69] S. Furui, "Unsupervised speaker adaptation method based on hierarchical clustering," in *Proc. ICASSP*, Glasgow, Scotland, May 1989, pp. 286–289.
[70] S. Cox, "Predictive speaker adaptation in speech recognition," *Comput. Speech Lang.*, vol. 9, no. 1, pp. 1–18, 1995.
[71] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation of multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 291–298, Apr. 1994.
[72] *Special Issue on Spoken Language Processing, Proc. IEEE*, Aug. 2000, vol. 88.
[73] M. G. Rahim and B. H. Juang, "Signal conditioning techniques for robust speech recognition," *IEEE Signal Processing Lett.*, vol. 3, pp. 107–109, Apr. 1996.
[74] ——, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 19–30, Jan. 1996.
[75] A. Sankar and C.-H. Lee, "Robust speech recognition based on stochastic matching," in *Proc. ICASSP*, vol. 1, Detroit, MI, 1995, pp. 121–124.
[76] M. J. F. Gales and S. J. Young, "Robust speech recognition in additive and convolutional noise using parallel model combination," *Computer Speech Language*, vol. 9, pp. 289–308, 1995.
[77] P. J. Moreno, B. Raj, E. Gouvea, and R. M. Stern, "Multivariate-Gaussian-based cepstral normalization for robust speech recognition," in *Proc. ICASSP*, vol. 1, Detroit, MI, 1995, pp. 137–140.
[78] L. R. Bahl, P. F. Brown, P. V. De Souza, and R. L. Mercer, "A tree-based statistical language model for natural language speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1001–1008, July 1989.
[79] R. Lau, R. Rosenfeld, and S. Roukos, "Trigger-based language models: A maximum entropy approach," in *Proc. ICASSP*, vol. 2, Minneapolis, MN, 1993, pp. 45–48.
[80] S. Deligne and F. Bimbot, "Language modeling by variable length sequences: Theoretical formulation and evaluation of multigrams," in *Proc. ICASSP*, Detroit, MI, 1995, pp. 169–172.
[81] P. F. Brown *et al.*, "Class-based N-gram models of natural language," *Computational Linguistics*, pp. 467–479, 1992.
[82] S. F. Chen and R. Rosenfeld, "A survey of smoothing techniques for ME models," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 37–50, Jan. 2000.
[83] J. R. Bellegarda, "Large vocabulary speech recognition with multispan statistical language models," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 76–84, Jan. 2000.
[84] R. Kuhn and R. De Mori, "A cache-based natural language model for speech recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, pp. 570–583, June 1990.
[85] M. J. Lee and L. F. Chien, "Automatic acquisition of phrasal knowledge for English–Chinese bi-lingual information retrieval," presented at ACM SIGIR, 1998.
[86] R. Haeb-Umbach and H. Ney, "Improvements in time-synchronous beam search for 10 000-word continuous speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 353–356, Apr. 1994.
[87] M. Bates, "Models of natural language understanding," in *Voice Communication between Humans and Machines*, D. Roe and J. Wilpon, Eds: Washington, DC National Academy, 1994, pp. 238–253.
[88] 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 Press, 1998, pp. 461–484.
[89] S. Seneff, "Robust parsing for spoken language systems," in *Proc. ICASSP*, 1992, pp. 189–192.
[90] R. De Mori, "Recognizing and using knowledge structures in dialogue systems," presented at the IEEE Workshop Automatic Speech Recognition and Understanding, 1999.
[91] P. Boda, "From stochastic recognition to understanding: An HMM-based approach," in *Proc. IEEE Workshop Automatic Speech Recognition and Understanding*, Santa Barbara, CA, 1997, pp. 57–64.
[92] W. Minker, "Design considerations for knowledge source representations of a stochastically based natural language understanding component," *Speech Commun.*, vol. 28, no. 2, pp. 141–154, 1999.
[93] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, pp. 39–41, 1995.
[94] R. J. Mammone *et al.*, "Robust speaker recognition—A feature-based approach," *IEEE Signal Processing Mag.*, vol. 13, pp. 58–71, Sept. 1996.
[95] A. E. Rosenberg *et al.*, "Speaker identification with user-selected password phrases," in *Proc. Eurospeech*, 1997, pp. 1371–1374.
[96] T. Nordstrom *et al.*, "A comparative study of speaker verification systems using the polycast database," in *Proc. ICSLP*, 1998, pp. 1359–1362.
[97] J. M. Pickett, J. Schroeter, C. Bickley, A. Syrdal, and D. Kewley-Port, "Speech technology," in *The Acoustics of Speech Communication*, J. M. Pickett, Ed. Boston, MA: Allyn & Bacon, 1998, ch. 17, pp. 324–342.
[98] 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.
[99] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava, "Speech and language technologies for audio indexing and retrieval," *Proc. IEEE*, vol. 88, pp. 1338–1353, Aug. 2000.
[100] J. R. Glass, "Challenges for spoken dialogue systems," presented at the IEEE Workshop Automatic Speech Recognition and Understanding, 1999.
[101] H. Aust, M. Oerder, F. Seide, and V. Steinbiss, "The Philips automatic train timetable information system," *Speech Commun.*, vol. 17, pp. 249–262, 1995.
[102] J. Glass, G. Flammia, D. Goodine, M. Phillips, J. Polifroni, S. Sakai, S. Seneff, and V. Zue, "Multilingual spoken-language understanding in the MIT voyager system," *Speech Commun.*, vol. 17, pp. 1–18, 1995.
[103] A. Gorin, G. Riccardi, and J. Wright, "How may I help you?," *Speech Commun.*, vol. 23, pp. 113–127, 1997.
[104] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. Hazen, and L. Hetherington, "Jupiter: A telephone-based conversational interface for weather information," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 85–96, Jan. 2000.
[105] V. Souvignier, A. Kellner, B. Rueber, H. Schramm, and F. Seide, "The thoughtful elephant: Strategies for spoken dialogue systems," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 1, pp. 51–62, Jan. 2000.
[106] S. Rosset, S. Bennacef, and L. Lamel, "Design strategies for spoken language dialog systems," in *Proc. Eurospeech*, 1999, pp. 1535–1538.
[107] M. Denecke and A. Waibel, "Dialogue strategies guiding users to their communicative goals," in *Proc. Eurospeech*, 1997, pp. 2227–2230.
[108] S. Sutton *et al.*, "Universal speech tools: The CSLU toolkit," in *Proc. ICSLP*, 1998, pp. 3221–3224.
[109] A. Waibel, P. Geutner, L. M. Tomokiyo, T. Schultz, and M. Woszczyna, "Multilinguality in speech and spoken language systems," *Proc. IEEE*, vol. 88, pp. 1297–1313, Aug. 2000.
[110] L.-S. Lee, "Voice dictation of Mandarin Chinese," *IEEE Signal Processing Mag.*, vol. 14, pp. 63–101, July 1997.
[111] ——, "Structural features of Chinese language—Why Chinese language processing is special and where we are," presented at the Int. Symp. Chinese Spoken Language Processing, Singapore, 1998.
[112] H. Ney *et al.*, "Algorithms for statistical translation of spoken languages," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 24–36, Jan. 2000.

Lin-shan Lee (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA.

He has been a Professor of Electrical Engineering and Computer Science at National Taiwan University since 1982, including being a department head of the university (1982–1987). He also holds a joint appointment as a Research Fellow of Academia Sinica, and was an institute director of Academia Sinica (1991–1997). His research interests include digital communications and Chinese spoken language processing. He developed several of the earliest versions of Chinese spoken language processing systems in the world, including a text-to-speech system, natural language analyzer, and dictation systems.

Dr. Lee has been a member of the Permanent Council of the International Conference on Spoken Language Processing (ICSLP). He was the Guest Editor for the Special Issue on Intelligent Signal Processing in Communications, IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, December 1994 and January 1995. He was the Vice President for International Affairs (1996–1997) and the Awards Committee chair (1998–1999) of the IEEE Communications Society.

Yumin Lee (Member, IEEE) was born in Taipei, Taiwan in 1968. He received the B.S. and M.S. degrees from National Taiwan University (NTU), Taipei, Taiwan, in 1989 and 1991, respectively, and the Ph.D. degree from Stanford University, Stanford, CA, in 1997, all in electrical engineering.

From 1989 to 1991, he was involved in the development of the world's first unlimited-vocabulary dictation system for Chinese Mandarin. From 1993 to 1997, he was a Research Assistant at Stanford University. From 1997 to 1999, he was affiliated with Motorola Labs, Motorola, Schaumburg, IL, where he worked on receiver performance analysis for and actively participated in the standardization activities of Enhanced Data Rates for GSM Evolution (EDGE). Since 1999, he has been an Assistant Professor in the Department of Electrical Engineering and Graduate Institute of Communication Engineering of NTU. His research interests include wireless and personal communications, communication theory, digital signal processing for communications, and speech processing.

Dr. Lee was a recipient of the Chinese Institute of Electrical Engineers (Taiwan) Graduate Student Thesis Award in 1991, and a Motorola UPR Grant Award from 1994 to 1997. He holds one U.S. patent.