T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

Author	B.-H. Juang, Sadaoki Furui
Journal/Book name	Proceedings of the IEEE, Vol. 88, No. 8, pp. 1142-1165
発行日 / Issue date	2000, 8
権利情報 / Copyright	(c)2000 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Powered by T2R2 (Science Tokyo Research Repository)

Automatic Recognition and Understanding of Spoken Language—A First Step Toward Natural Human–Machine Communication

BIING-HWANG JUANG, FELLOW, IEEE, AND SADAOKI FURUI, FELLOW, IEEE

Invited Paper

The promise of a powerful computing device to help people in productivity as well as in recreation can only be realized with proper human–machine communication. Automatic recognition and understanding of spoken language is the first and probably the most important step toward natural human–machine interaction. Research in this fascinating field in the past few decades has produced remarkable results, leading to many exciting expectations as well as new challenges. In this paper, we summarize the development of the spoken language technology from both a vertical (the chronology) and a horizontal (the spectrum of technical approaches) perspective. We highlight the introduction of statistical methods in dealing with language-related problems as it represents a paradigm shift in the research field of spoken language processing. Statistical methods are designed to allow the machine to learn, directly from data, structure regularities in the speech signal for the purpose of automatic speech recognition and understanding. Today, research results in spoken language processing have led to a number of successful applications, ranging from dictation software for personal computers and telephone-call processing systems for automatic call routing to automatic subcaptioning for television broadcast. We analyze the technical successes that support these applications. Along with an assessment of the state-of-the-art in this broad technical field, we also discuss the limitations of the current technology and point out challenges that are ahead of us. We hope that through this paper an accurate overview of the spoken language technology can be presented as the basis to inspire future advances.

***Keywords*—Acoustic modeling, acoustic-phonetics, articulation, automatic recognition and understanding, Bayes risk, cepstral distance, continuous speech recognition, detection-based approach, dialogue systems, discriminative training, dynamic programming, finite state machine, forward–backward algorithm, generalized phone models, grammar, hidden Markov models, human–machine communication, isolated word recognition, language modeling, language structure, linear prediction, maximum a posteriori, maximum-likelihood estimation, noise, perplexity, probability distribution of speech, pronunciation modeling, robustness, search algorithms, short-time spectral analysis, signal analysis, speech dictation, speech distortion, speech representations, spoken language processing technology, statistical language processing, statistical pattern recognition.**

Manuscript received October 30, 1999; revised March 20, 2000.

B.-H. Juang is with Lucent Technologies, Inc., Murray Hill, NJ 07974 USA.

S. Furui is with the Tokyo Institute of Technology, Tokyo, 152–8552 Japan.

Publisher Item Identifier S 0018-9219(00)08096-8.

I. Introduction

Speech is the primary, and the most convenient, means of communication between people. Whether it is due to the technological curiosity to build machines to mimic humans or the desire to automate work with machines, research in machine recognition of human speech, as the first step toward natural human–machine communication, has attracted much enthusiasm over the past four decades. The advent of powerful computing devices further gives hope to this relentless pursuit, particularly in the past few years. While we are still far from having a machine that converses with a human like a human, many important scientific advances have taken place, bringing us closer to the "Holy Grail" of automatic speech recognition and understanding by machine. To gain an appreciation of the amount of progress, the scope, and the associated technical difficulties in spoken language processing (SLP), it is worthwhile to briefly review several milestones in the field. Such a review also paves the way to a better understanding of the currently prevalent technical framework, which forms the foundation of many speech-recognition products and services used in real-world, albeit limited, applications.

A. Historical Perspective of Research in Human–Machine Communication by Speech

Early attempts to design systems for automatic speech recognition were mostly guided by the theory of acoustic-phonetics, which describes the elements of

0018–9219/00$10.00 © 2000 IEEE

speech and how they are realized to form a spoken language. In 1952, Davis *et al.* of Bell Laboratories built a system for isolated digit recognition for a single speaker [1], using the spectral resonances during vowel regions of each digit. In 1956, Olson and Belar of RCA Laboratories tried to recognize ten syllables of a single talker [2]. At MIT Lincoln Laboratory, Forge and Forge built a speaker-independent ten-vowel recognizer in 1959, using time-varying estimates of the vocal tract resonance [3]. Later, in the 1960s, with emphasis on building a special hardware, several Japanese laboratories also demonstrated their progress. Most notable among them were the vowel recognizer of Suzuki and Nakata of the Radio Research Lab in Tokyo [4], the phoneme recognizer of Sakai and Doshita of Kyoto University (noting the use of a speech segmenter to allow analysis and recognition of speech in different portions of the signal) [5], and the digit recognizer of NEC Laboratories [6].

One significant remark to be made is the year 1959 when Fry and Denes, at University College in England, attempted a phoneme recognizer to recognize four vowels and nine consonants [7], [53]. They incorporated statistical information about allowable phoneme sequences in English to enhance the overall phoneme recognition accuracy for words consisting of two or more phonemes. This perhaps marked the first use of statistical syntax in automatic speech recognition.

The work of Martin's team at RCA Laboratories and that of Vintsyuk in the Soviet Union in the 1960s have particularly important implications on the research and development of automatic speech recognition. Martin recognized the need to deal with the nonuniformity of time-scale in speech events and suggested realistic solutions, including detection of utterance endpoints, which greatly enhanced the reliability of the recognizer performance [8]. Vintsyuk proposed the use of dynamic programming for time-alignment between two utterances in order to derive a meaningful matching score [9]. Although his work was largely unknown to the West then, it appears to have preceded that of Sakoe and Chiba [10], as well as others who proposed more formal methods in speech pattern matching, generally known as dynamic time warping. Since the late 1970s, dynamic programming, in numerous variant forms, has become an indispensable technique in the pattern-matching approach to automatic speech recognition.

Two broad directions in speech-recognition research started to take shape in the 1970s, with IBM and Bell Laboratories essentially representing the two different schools of thought in terms of the intermediate goals of speech communication between human and machine. IBM's effort, led by Jelinek, was aiming at a voice-activated typewriter, the main function of which was to convert spoken sentences into a sequence of letters and words that could be shown on a display or typed on paper [11], [12]. The system was mostly speaker-dependent (i.e., the typewriter was to be trained to listen to its owner or primary user) and the technical focus was on the structure of language. In the approach, the language structure is represented by a probabilistic model, which describes how likely a sequence of linguistic symbols (e.g., phones or words) can appear in the speech signal. This type of task is often referred to as "transcription." At Bell Laboratories, the goal was to provide telecommunication services to the public, such as voice dialing, and command and control for automation of phone calls. For most of the applications of this kind, the system is expected to work well for a vast population of talkers, independent of the identity of the talker. The focus at Bell Laboratories was then in the design of a speaker-independent system that could deal with the acoustic variability intrinsic in the speech signals coming from many different talkers, often with notably different regional accents [13]. Research to understand and to harness the acoustic variability manifests itself in the study of spectral distance measures (see, e.g., [14] and [15]) and clustering techniques [16]. Also of importance in the Bell Laboratories' approach to the problem is the concept of keyword spotting as a primitive form of speech understanding [17]. Keyword spotting attempts to detect prescribed words or phrases of particular significance, while neglecting those nonessential portions of the utterance. This is owing to the need to accommodate talkers who often prefer to speak natural sentences rather than rigid command words. These two approaches had a profound influence in the evolution of the human–machine speech communication technology in the past two decades. One common theme between these modern efforts, despite the differences, is that mathematical formalism and rigor start to emerge as a distinct and important aspect of speech research.

Another achievement in parallel to the above developments was the work of Reddy at Carnegie-Mellon University, who first advocated dynamic phoneme tracking for continuous speech recognition [18] and later proposed a knowledge integration approach to speech recognition and understanding in the context of artificial intelligence research [19].

While the difference in goals led to different realizations of the technology in various applications, the rapid development of statistical methods in the 1980s, namely the hidden Markov model (HMM) framework [19]–[21], had caused a certain degree of convergence in the system design. Today, most of the systems in use are based on the statistical framework and results developed in the 1980s, with additional improvements in the 1990s. We shall elaborate in the following a communication-theoretic framework, which supported the development of the fundamental method.

B. Communication-Theoretic Framework

Speech communication involves sensory as well as cognitive behaviors. Traditionally, the speech communication chain comprises four stages: detection of acoustic-phonetic cues to form words, syntactic and grammatical analysis for parsing of sentences and for error correction, semantic determination and disambiguation, and pragmatic inference with additional prosodic cues and interpretation of message intent. The ultimate machine that can converse with a human would need all the knowledge to perform those four stages of the speech communication chain. It involves understanding of the context (the subject domain as well as the mood and the ambient) of the conversation. No attempt has yet reached

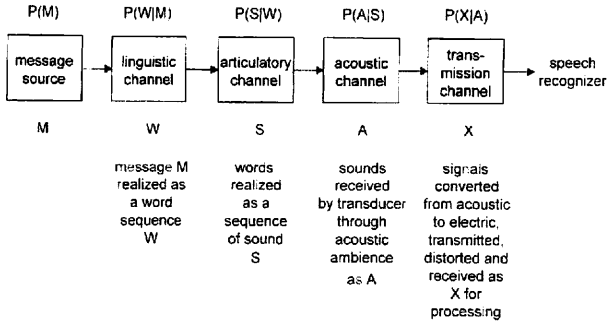


Fig. 1. Communication-theoretic formulation of the speech production chain.

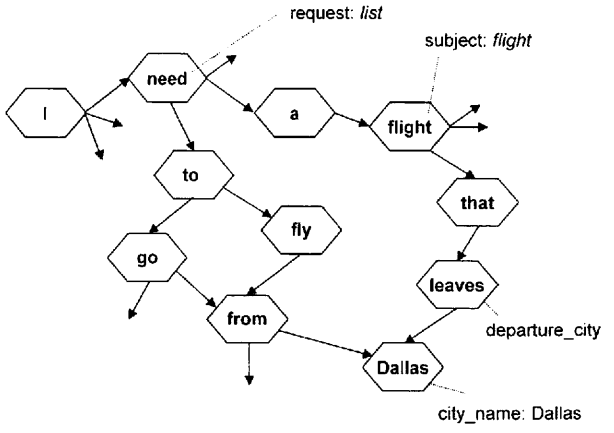


Fig. 2. Finite state network.

such a level of complexity. While academic research following every step of this classical formulation can be contemplated, speech technologies developed so far often take a more simplified and restricted view. The machine that we have attempted to design so far is, almost without exception, limited to the simple task of converting a speech signal into a word sequence and then determining, from the word sequence, the meaning that is "understandable." Here, the set of understandable messages is finite in number, each being associated with a particular action (e.g., route a call to a proper destination, or issue a buy order for a particular stock). In this limited sense of speech communication, the focus is detection and recognition rather than inference and generation.

Following this limited goal of human–machine communication, a concrete and yet convenient way to describe the speech generation/production chain is shown in Fig. 1, which depicts the basis of a communication-theoretic approach to automatic speech recognition and understanding. In this formulation, a message source decides to convey an intended message $\boldsymbol{M}$, which is realized as a word sequence $\boldsymbol{W}$ through a linguistic channel, specified by a probability measure $P(\boldsymbol{W}|\boldsymbol{M})$. The linguistic channel is probabilistic as there are many ways to express the same message, some more likely than others. For example, Fig. 2 (after [23]) shows partially a finite state network of numerous expressions, all leading to the same semantic message: *the user needs information about flights that leave Dallas*. The word sequence $\boldsymbol{W}$ then gets realized, through the articulatory

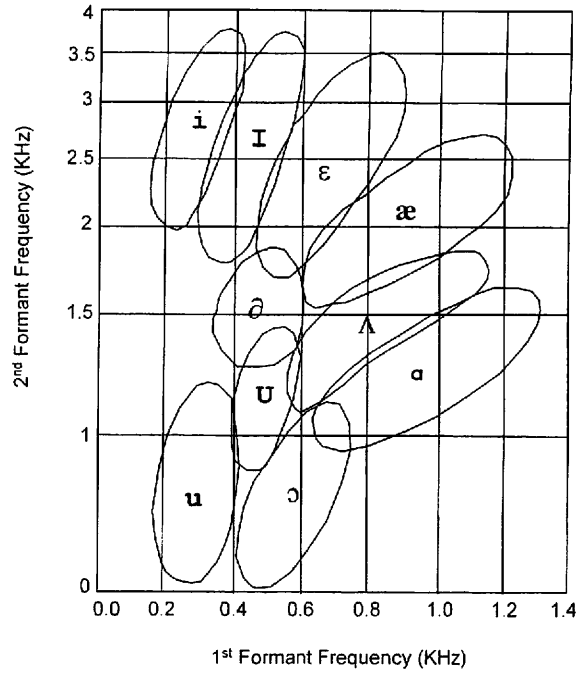


Fig. 3. Distribution of vowels in F1–F2 plane.

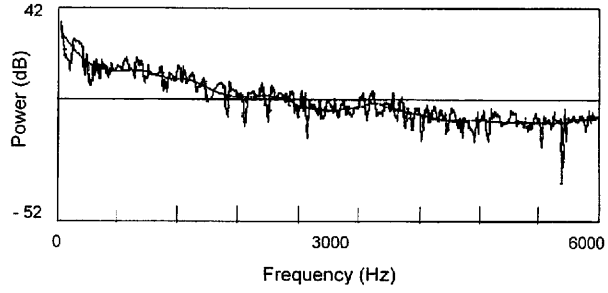


Fig. 4. Typical noise spectrum in an office with a Sun workstation running.

channel, as a sequence of sounds $\boldsymbol{S}$, which may be expressed in terms of phonemes as they are considered the fundamental elements of speech. Again, the articulatory channel $P(\boldsymbol{S}|\boldsymbol{W})$ introduces variability because no one talker can repeat exactly the same waveform even uttering the same word, and no two talkers are alike in terms of the configuration of their articulatory apparatus. Fig. 3 shows a well-known vowel triangle in the plane of the first and the second resonant frequency produced by a population of talkers [24]. The spread signifies the extent of variability in the resonant frequencies. The sequence of sounds $\boldsymbol{S}$ is radiated from the mouth of the talker, propagates in acoustic waves through the room, is convolved with the room acoustic response and mixed with the acoustic ambient, and then reaches the microphone as the acoustic input $\boldsymbol{A}$ We label this process "acoustic channel." The acoustic ambient in various rooms can be quite different. Fig. 4 shows the power spectrum of a typical acoustic background noise in a personal office with a computer running. Fig. 5 shows an example of the impulse response of a room. These ambient conditions can vary tremendously from one to another. The acoustic channel is characterized by a probabilistic model $P(\boldsymbol{A}|\boldsymbol{S})$. The

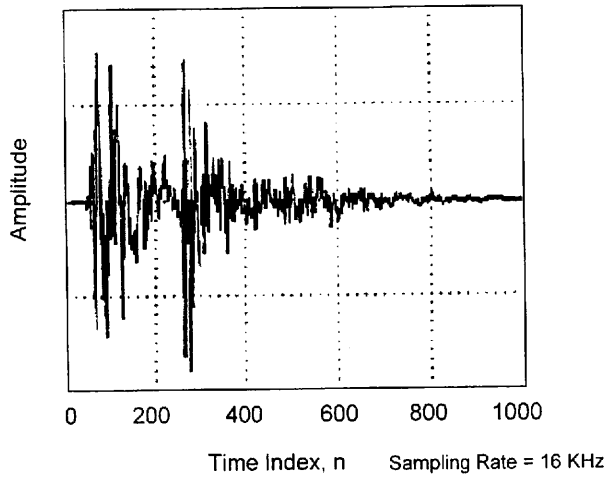


Fig. 5. Impulse response in a typical conference room.

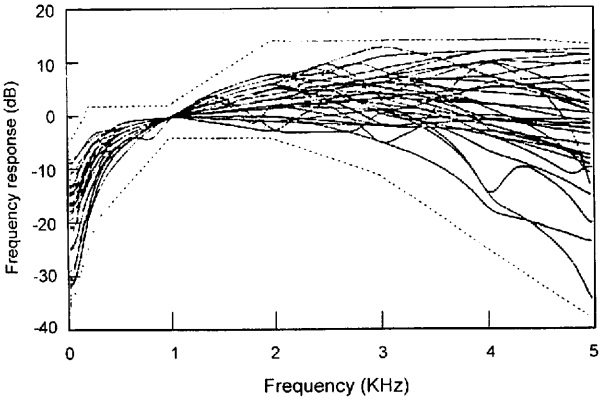


Fig. 6. Diversity of transducer characteristics in telephone set [25].

acoustic input signal $\boldsymbol{A}$ finally is converted by a microphone into an electric signal, which then propagates through the transmission route (cables, wires, or the telephone network) and becomes $\boldsymbol{X}$ when it is received by the recognition and understanding system. Microphone responses can also vary substantially; for example, as shown in Fig. 6 (after [25]), the frequency response of a telephone set can vary by as much as 40 dB at 4 kHz. This tremendous variation can make the comparison of speech patterns unreliable if not properly equalized. This last stage of uncertainty (transducer variation and transmission distortion) is called the "transmission channel," characterized by a probability model $P(\boldsymbol{X}|\boldsymbol{A})$. An automatic speech recognition and understanding system tries to reverse the process to recover $\boldsymbol{M}$.

In this approach, every channel represents one class of uncertainty or variation and is individually characterized by a probability distribution. The message source may also have a probabilistic prior $P(\boldsymbol{M})$, as some messages may be more likely to appear than others. (Note that $\boldsymbol{X}$, $\boldsymbol{A}$, $\boldsymbol{S}$, $\boldsymbol{W}$, and $\boldsymbol{M}$ are considered random events in the context of probability theory.) The availability of knowledge of these uncertainties dictates how well a system can communicate with people. This formulation also provides a unified framework for developing necessary technologies and systems in many practical applications. In the following sections, we will elaborate recent technological developments and justification in systems designs using this model of speech generation and communication.

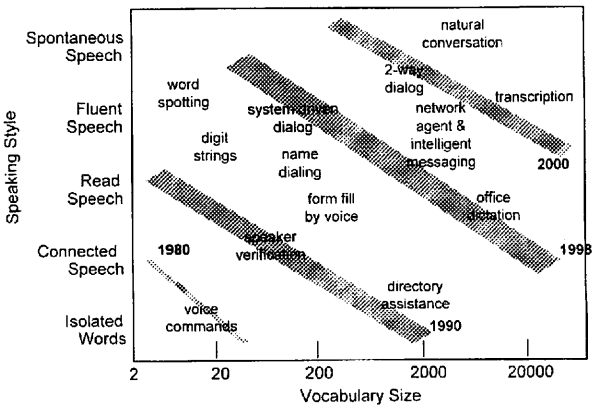


Fig. 7. Progress of spoken language technology along the dimensions of vocabulary size and speaking styles.

C. State-of-the-Art and Current Status in Commercialization of the Technology

Research efforts in automatic speech recognition and understanding in the past few decades have made viable a number of application tasks involving human–machine interaction via speech. As a result, commercialization of spoken language processing technologies is currently experiencing a strong push from many laboratories, companies, and research institutes. These applications can be broadly divided into telecommunication and nontelecommunication areas. In nontelecommunication applications, use of speech recognition systems mostly takes the form of personal computer (PC) software with automatic dictation as its primary use. Leaders in this category of software include IBM, Dragon Systems, L&H, and Philips. Although these software packages have yet to be improved for general speech-to-text conversion purposes (e.g., converting voice-mails into text), they have received positive reviews from professional groups such as radiologists and lawyers in their specialized field of applications. In the telecommunication application arena, most of the applications are in automating calls, which can be astronomical in terms of the number of sessions and number of users. It requires a true speaker-independent system. In a successful deployment of a call-processing application (AT&T's VRCP), it has been reported that a five-word keyword spotting system automates billions of calls every year resulting in savings in operating cost in hundreds of million of dollars [26]. Other service oriented applications of automatic speech recognition and understanding systems include on-line stock transactions and credit-card account services and management.

The state-of-the-art in automatic speech recognition can be addressed in several ways. We present two here. Fig. 7 illustrates the progress of speech recognition and understanding technology according to generic application areas, ranging from isolated word/command recognition to natural conversation between human and machine [27]. The complexity of these generic application areas is characterized along two di-

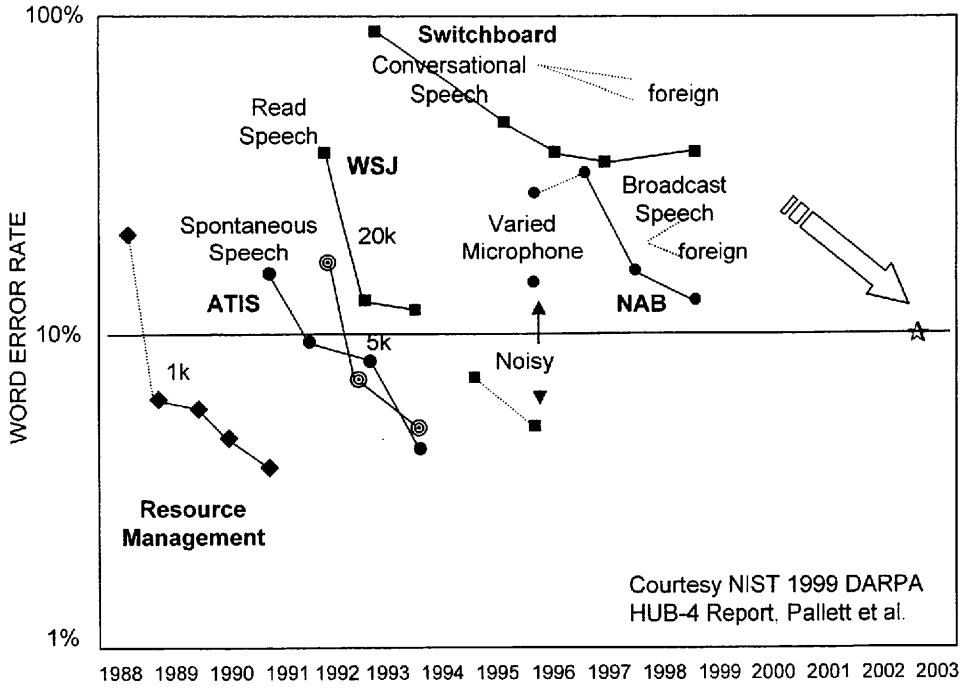


Fig. 8. Benchmarks of ASR performance in word error rates in DARPA-sponsored tasks [28].

mensions: the size of the vocabulary and the speaking style. It should be obvious that the larger the vocabulary, the more difficult the application task. Similarly, the degree of constraints in the speaking style has a very direct influence on the complexity of the application: a free conversation full of slurring and extraneous sounds such as "uh," "um," and partial words is far more difficult than words spoken in a rigidly discrete manner. Thus, the difficulty of an application grows from the lower left corner to the upper right corner in Fig. 7. The fuzzy diagonal lines demarcate the applications that can and cannot be supported by the technology for viable deployment in the corresponding time frame.

Another benchmark of the technology is shown in Fig. 8, which represents the chronological progress in various tasks sponsored by the DARPA program [28]. Plotted in the figure are the smallest word error rates achieved by the participants in various DARPA evaluation contests during the past decade. These error rates are often considered optimistic due to the fact that the data are all prerecorded, particularly for those "read speech" (i.e., speech signals from a talker reading a prepared text), and lack the often unexpected variability that exists in most real speech interactions. As is obvious from the chart, the word error rate for conversational speech remains extremely high, at close to 40% level. Therefore, current spoken language technologies are, in general, deemed inadequate as a machine intended to converse with a human. The technology evolution, thus, continues.

D. Purpose and Outline of this Paper

This paper is intended to be a brief summary of the technology development in the field of spoken language processing, particularly in automatic speech recognition and understanding, during the last two decades. While omission and personal biases are inevitable, we hope to objectively present a reasonable portrait of the technological advances for the purpose of inspiring more in the future.

This paper is organized as follows. We state the problem and the basic formulation of the approaches to the problem in the next section. A contrast between the acoustic-phonetic approach and the data-driven statistical approach is given so as to induce the possibility of cross-fertilization, because these two dominant methodologies can exist in a complementary manner. We discuss the possibility of various goals and performance criteria in the design of a speech recognition and understanding system. In Section III, we present the fundamental system architecture and elaborate its technological components that have become the focal points of recent spoken language research. Having discussed the technical details of the system components, we address and present in Section IV a number of research issues that await further research. We conclude this paper in Section V.

II. Problem Statements and Foundations of Speech Recognition and Understanding

Spoken language processing encompasses a broad range of technical challenges, including recognition of words and phrases in the speech signal, extraction of keywords or key phrases in the utterance, and understanding of the spoken utterance for the machine to take actions. Conversation between people can take many different forms, many of which may be beyond the scope of the current scientific interest. For example, a casual conversation between two people can drift over an unbounded domain with no end result anticipated. We will not address this category of scenarios. We will, however, assume that the common goal in speech recognition and understanding is to identify an important message, out of a

finite set of possibilities, conveyed in the spoken utterance. In order to achieve this goal, the technology may choose to identify intermediate results such as phonemes, words, or sentences. We also often choose to judge the performance of such a machine by its error probability (or error rate, as an empirical estimate of the error probability) in making a recognition decision upon a test.

Recognition is a capability that can be addressed in at least two ways in our daily life. One is to be able to identify a particular object of interest, say, a particular model of car made by a manufacturer or a song from a particular artist. The other is to be able to recognize an object from a finite and closed set of possibilities, much like a multiple choice test, the average score (percent correct) of which is often used to judge the subject's performance. In the former scenario, the subject under test evaluation may or may not know the entire universe that can cause confusion to the correct answer so long as it has enough evidence to make a determination. If it makes a correct identification (e.g., naming the car model or the song of the artist), we say it *can* recognize; otherwise, it cannot. The latter situation is different in that the subject knows all about the task and its scope (choosing one out of a finite number of answers), but due to the uncertainty in the observed evidence, its answer may be at times incorrect. The performance in the latter case is then usually measured in terms of the recognition accuracy (percent correct in answers to a large set of trials or tests) or error rate (percent incorrect in answers). The former case is the basis of many classical speech-recognition research efforts, while the latter case gives rise to the statistical approach to the pattern-recognition problem that finds widespread engineering applications in recent years.

A. Basics of Linguistics and Acoustic-Phonetics Related to SLP

Most of the classical speech-recognition research was based on the identification paradigm as discussed above. It requires extensive understanding of the properties of the object (i.e., the speech sound). It, thus, depends on and makes use of, almost exclusively, the acoustic-phonetic theory, which aims at building a framework for understanding speech by a human.

Phoneticians and linguists decompose a spoken language into elements of linguistically distinctive sounds—the phonemes. The number of phonemes in a language is often a matter of judgment and is not invariant to different linguists. Phonemes are determined and taxonomically classified according to their corresponding articulatory configurations. For example, a vowel is produced by exciting a vocal tract of an essentially fixed shape with quasi-periodic pulses of air, caused by the vibration of the vocal cords. Front vowels (/i/, /I/, /e/, and /æ/) are vowels produced with a tongue hump in the front portion of the vocal tract. Other phoneme categories include diphthongs, semivowels, nasals, stops, fricatives, affricates, and whisper. As in many classical studies, the taxonomy was established for a systematic investigation of the properties of the "element" of speech sounds. Such properties of sounds are often referred to as acoustic-phonetic features. An alternative way to classify the phonemes is to use the broad phonetic class according to key acoustic-phonetic feature dimensions.

The long history of acoustic-phonetic studies has produced a fairly extensive understanding of the properties of phonemes, particularly in terms of their general behavior. The acoustic-phonetic knowledge scientists were able to accumulate has guided the main development of spoken language processing technologies in the past. The knowledge is, however, insufficient when it comes to dealing with variability in speech. First, due to the limitation of computing and recording tools, previous studies of acoustic-phonetics tend to focus more on "typical" and "standard" behaviors. Second, most of the speech materials for linguistic studies in the past were recorded under well-controlled (and clean) conditions; rarely had noisy or distorted speech been extensively investigated. Behaviors of "found" speech—that exists ubiquitously—are, thus, less well known. The scope of understanding in the sound variability is, thus, often limited. Recent investigations in automatic recognition and understanding of speech differ from the classical acoustic-phonetic approach in the requirement for proper handling of the extensive variability exhibited in the speech sounds produced by lay people in their everyday life (as opposed to a professional narrator speaking in a quiet studio or sound booth). The ability to deal with the statistical behavior of spoken utterances is imperative due to the prescribed *average* performance criterion, as will be discussed shortly.

B. Statistical Pattern Recognition Formulation—A Data-Driven Approach

The formulation of statistical pattern recognition has its root in Bayes' decision theory. Let $\boldsymbol{X}$ be a random observation from an information source, consisting of M classes of event. A classifier's job is to correctly classify each $\boldsymbol{X}$ into one of the M classes. (Here, we use the terms *classifier* and *recognizer* interchangeably because we have defined the problem as identifying an unknown observation as one of M classes of event.) We denote these classes by C_i, $i = 1, 2, \ldots, M$. Let $P(\boldsymbol{X}, C_i)$ be the joint probability distribution of $\boldsymbol{X}$ and C_i, a quantity that is assumed to be known to the designer of the classifier. In other words, the designer has full knowledge of the random nature of the source.

To measure the performance of the classifier, we further define for every class pair (i, j) a cost or loss function c_{ij}, which signifies the cost of classifying (or recognizing) a class i observation into a class j event. The loss function is generally nonnegative, with $c_{ii} = 0$ representing a correct classification.

Given an arbitrary observation $\boldsymbol{X}$, a conditional loss for classifying $\boldsymbol{X}$ into a class i event can be defined as [29]

$$R(C_i|\boldsymbol{X}) = \sum_{j=1}^{M} c_{ij} P(C_j|\boldsymbol{X}) \mathbf{1}(\boldsymbol{X} \in C_j) \qquad (1)$$

where $P(C_j|\boldsymbol{X})$ is the *a posteriori* probability and $1(\bullet)$ is the indicator function. This leads to a reasonable performance measure for the classifier, the expected loss, defined as

$$L = \int R(C(\boldsymbol{X})|\boldsymbol{X})p(\boldsymbol{X})\,d\boldsymbol{X} \tag{2}$$

where $C(\boldsymbol{X})$ represents the classifier's decision, assuming one of the M "values," $C_1, C_2, \ldots, C_M$ based on a random observation $\boldsymbol{X}$ drawn from a probability distribution $P(\boldsymbol{X})$. The decision function $C(\boldsymbol{X})$ depends on the classifier design. Obviously, if the classifier is so designed that for every $\boldsymbol{X}$

$$R(C(\boldsymbol{X})|\boldsymbol{X}) = \min_i R(C_i|\boldsymbol{X}) \tag{3}$$

the expected loss in (2) will be minimized. For speech recognition, the loss function e_{ij} is usually chosen to be the zero–one loss function defined by

$$e_{ij} = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases} \quad i, j = 1, 2, \ldots, M \tag{4}$$

which assigns no loss to a correct classification and a unit loss to any error, regardless of the class. With this type of loss function, the expected loss L is, thus, the error probability of classification (or recognition). The conditional loss becomes

$$\begin{aligned} R(C_i|\boldsymbol{X}) &= \sum_{j \neq i} P(C_j|\boldsymbol{X}) \\ &= 1 - P(C_i|\boldsymbol{X}). \end{aligned} \tag{5}$$

The optimal classifier that achieves minimum L is, thus, the one that implements the following:

$$C(\boldsymbol{X}) = C_i \quad \text{if } P(C_i|\boldsymbol{X}) = \max_j P(C_j|\boldsymbol{X}). \tag{6}$$

In other words, for minimum error rate classification, the classifier employs the decision rule of (6), which is called the "maximum *a posteriori*" (MAP) decision. The minimum error rate achieved by the MAP decision is called the "Bayes risk" [29].

The required knowledge for an optimal classification decision is, thus, the *a posteriori* probabilities for implementing the MAP rule. These probabilities, however, are not given in practice and have to be estimated from a training set of observations with known class labels. The Bayes decision theory, thus, effectively transforms the classifier design problem into a distribution estimation problem. The significance of this approach is that the knowledge required in the system design can be directly learned from the data, without intensive deduction from human experts. Thus, the basis of the statistical approach to pattern recognition can be stated as follows: collect and label (with certainty) a set of observations (design sample) $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_I\}$, and estimate the *a posteriori* probability distribution $P(C_i|\boldsymbol{X})$, $i = 1, 2, \ldots, M$ to implement the maximum *a posteriori* decision to achieve the (minimum) Bayes risk. The *a posteriori* probability $P(C_i|\boldsymbol{X})$ can be rewritten as

$$P(C_i|\boldsymbol{X}) = P(\boldsymbol{X}|C_i)P(C_i)/P(\boldsymbol{X}). \tag{7}$$

Since $P(\boldsymbol{X})$ is not a function of the class index, and, thus, has no effect on the MAP decision, the needed probabilistic knowledge can be represented by the class prior $P(C_i)$ and the conditional probability $P(\boldsymbol{X}|C_i)$.

In the context of speech recognition and understanding, a class identity may be associated with a word in the vocabulary, or a sequence of words or phonemes that is designated as a unitary linguistic event. The conditional probability $P(\boldsymbol{X}|C_i)$ characterizes the randomness in realizing a class C_i event in the signal $\boldsymbol{X}$.

Despite the advantage of automatic learning from data, there are several issues associated with this classical approach. First, the distributions usually have to be parameterized in order for them to be practically useful for the implementation of the MAP rule. This is particularly necessary in the case of the conditional probability $P(\boldsymbol{X}|C_i)$ when $\boldsymbol{X}$ is a continuously valued vector quantity. It is parameterized as $P_\Lambda(\boldsymbol{X}|C_i)$ and is called the acoustic model. Correspondingly, the prior $P(C_i)$, which defines the distribution of the linguistic event, is represented by a measure called the language model, particularly when the classes are associated with the words in the vocabulary. The classifier designer, therefore, has to determine the right parametric form of the distributions. For most real-world problems, this is a difficult task. Our choice of the distribution form is often limited by the mathematical tractability of the particular distribution function and is very likely to be inconsistent with the actual data distribution. This means the true MAP decision can rarely be implemented, and the minimum Bayes risk generally remains an unachievable lower bound. Second, given a parameterized distribution form, the unknown parameters defining the distribution have to be estimated from the training data. A good parameter estimation method is, therefore, necessary. The estimation method has to be able to produce consistent parameter values. Third, the approach requires a training set of known examples. To reliably estimate the parameters, the training set needs to be of sufficient size. Usually, the more the training data is provided, the better the parameter estimate is. The difficulty, nevertheless, is that data collection and labeling is a labor-intensive and resource-demanding process, particularly for speech-recognition applications. When the amount of training data is limited, the quality of the estimated distribution parameters cannot be guaranteed.

These three basic issues point out a fundamental fact in the statistical pattern-recognition approach: that is, despite the conceptual optimality of the Bayes decision theory and its applications to pattern recognition, it cannot be accomplished because practical "MAP" decisions in speech recognition are not true MAP decisions. Fortunately, practical procedures to address these issues do exist, but this prior understanding is necessary in our discussions below.

C. Speech Variability

Speech variability refers to the uncertainty a speech recognition and understanding system would observe in the signal it receives from the speaker. As discussed above, the signal is

received via a speech production "channel," which can be decomposed into several stages, each causing a certain type of ambiguity to the original message during its realization. The key idea behind the statistical approach is to obtain an accurate characterization of the variation inherent in the speech signal in order to be able to implement the MAP decision rule as closely as possible. This requires first an understanding of the variability of speech, in terms of its source as well as its behavior.

When the message is translated into a sequence of words to form an acceptable sentence or phrase, or simply an "understandable utterance," there exist many possibilities, depending on many factors such as the talker's mood, the circumstance of the conversation, and so on. This kind of linguistic variability is probably the least known in terms of its statistical behavior. Only in extremely restricted tasks is there some database for analyzing this effect. Even then, the circumstantial influence is often untractable.

When words are pronounced as a sequence of phonemes, the variability comes from the fact that talkers may differ in their lexical habit, some due to regional accents and some due to education and upbringing. Pronunciation variation may also exist beyond the word boundary; the context may modify the pronunciation differently for different talkers. Representation of a "spoken" lexicon is a major issue in machine recognition and understanding of a spoken language.

Much of the speech variability comes from articulatory variation. Various speakers have different vocal tract configurations, shapes, and lengths. Very rarely, if not impossibly, would different talkers be able to produce acoustically identical sounds. Articulation requires motor control of the articulatory apparatus and can hardly be repeated even for the same speaker. Attempts in the past to model the articulatory variation mostly focus on talker normalization, aiming at transforming the speech spectrum to reflect the (static) parametric variation in the vocal tract apparatus (e.g., the vocal tract length [30]). No attempt exists, to the best of our knowledge, to normalize the variation in motor control of the articulatory apparatus.

Other sources of variability come from the ambient and the transmission channel. Background noise is ubiquitous and may display a vast range of characteristics and levels. Noise can take the shape of a door slam, a constant fan noise from a machine, the rotor noise of a chopper, or simply a background conversation. The most difficult to characterize is probably the nonstationary noise that resembles the speech itself. Distortions in transmission can come from the microphone arrangement (different types of microphone, or the telephone handset, which has a rather wide range of allowable deviation in frequency response), the transmission equipment (e.g., a speech coder in digital telephony or in a cellular phone network), or convolution with the room acoustic response. The reverberation from an echoic room poses as a major technical difficulty in achieving high recognition accuracy. As the use of hands-free speakerphones becomes more prevalent, the problem of reverberation will demand closer attention.

Having the insight of the source and the range of variability, one still needs to be able to translate it into a model, or a structural representation, for incorporation into the speech distribution. Much research in the past few years has been concentrating on this aspect.

D. From Recognition to Inference—Machine Intelligence for Communication

We consider speech understanding as capturing a particular notion that is embedded in the spoken utterance. The notion may appear in the form of a key word, a key phrase, or an expression. To understand the intended notion may require inference, involving the knowledge of the context in which the utterance appears, the pragmatic aspect of the expression, or the atmosphere that leads to the conversation. In other words, to go from automatic speech recognition to understanding may incur the need of a machine with more intelligence than the current speech systems. We will be directing our attention to only identification of a notion from a finite set of possibilities rather than machine intelligence research. It is possible, however, to design a system to have a dialogue with a human, emulating an intelligent machine, by a proper design of the dialogue flow involving simple identification of limited notions.

III. Technology Components of Automatic Speech Recognition and Understanding

In this section, we discuss a typical system and its technological components for automatic speech recognition and understanding. This represents a common architecture supporting most of today's engineering implementations of a spoken language processing system.

A. Simplified Framework for Speech-to-Text Conversion

As mentioned at the beginning of the article, there are two broad approaches to spoken language understanding. One assumes that each utterance comprises a sequence of linguistically meaningful and structured words, and the first step toward the goal of understanding is to convert the spoken acoustic signal into the word sequence as accurately as possible. Understanding of the meaning and the intention of the spoken message would follow based on the recognized sequence of words. The other approach makes no explicit assumption on the linguistic structure of the utterance; rather, it only attempts to deal with situations in which the intended message is always expressed in certain key words or phrases. (We shall come back and address the issues in speech understanding in later sections.) The first approach is particularly appropriate in the design of a voice-activated typewriter [11]. The task is sometimes referred to as "speech-to-text conversion" or "word decoding" in continuous speech recognition.

Formulation of word decoding is essentially based on the same Bayes decision theory discussed in Section II-B, except that the observation sequence may consist of more than one class of observations. Let $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$ be an (unknown) observation that may be an acoustic realization of

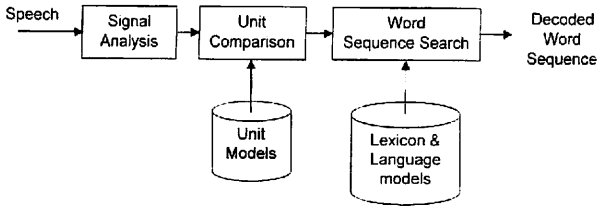


Fig. 9. Fundamental block diagram for an automatic speech recognition and understanding system based on the simplified decoding strategy.

a sequence of words $W = (w_1, w_2, \ldots, w_S)$, where each $w_i \in V$, the vocabulary. The speech recognizer attempts to implement the maximum *a posteriori* rule to find W^o

$$W^o = \arg\max_W P_A(W|\boldsymbol{X}) = \arg\max_W P_A(\boldsymbol{X}|W)P_A(W)/P_A(\boldsymbol{X}). \quad (8)$$

The key quantities affecting the decision are of course $P_A(\boldsymbol{X}|W)$ and $P_A(W)$ since $P_A(\boldsymbol{X})$ is not involved in the optimization process. As before, $P_A(\boldsymbol{X}|W)$ is related to the probabilistic realization of the word sequence and is called the acoustic model. The other quantity, $P_A(W)$, defines the probabilistic relationship that exists among words when they appear in sequence and is usually called the language model. Fig. 9 is a block diagram depicting the modules employed in a prevalent speech recognition system according to the above-simplified formulation. The received speech signal first goes through a signal analysis module in which the speech waveform is translated into a speech pattern representation, consisting of a sequence of feature vectors. The speech pattern is then compared with the reference patterns pretrained and stored with class identities. Such a comparison can involve several layers of processing, from a distance or likelihood calculation between two vectors, to a search procedure for detecting the presence of higher level units (i.e., phonemes, words, and possibly phrases and sentences). The decoded sequence of linguistic symbols is then subject to parsing and interpretation to infer the message intended in the utterance.

One should note that this simplified formulation of the recognition problem obviously lumps all the potential variation in the signal together and does not individually deal with each source of variability in an explicit manner. For an improved performance of the system, there may be a need to design the system to cope with each variability source separately in a multistage fashion as outlined in the previously discussed communication-theoretic framework.

B. Signal Analysis and Representation

Signal analysis is the first step in every automatic speech recognition and understanding system. The aim of signal analysis is to obtain the salient feature in the speech waveform that is critical to the recognition or identification of the unknown linguistic event. Speech waveform has various

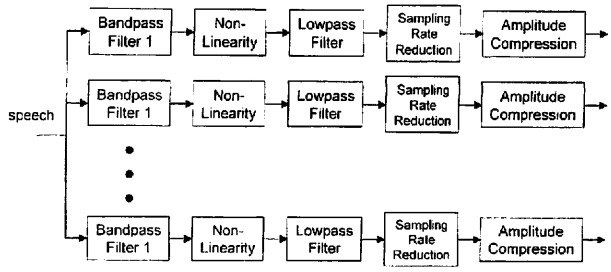


Fig. 10. Spectral envelope estimation using filter-bank.

kinds of features: for example, some pertain to the gender of the speaker, some relate to the quality of the sound, and some carry the necessary information for the intended message from the speaker. Many feature parameters can be directly measured or estimated from the acoustic waveform, while others such as gender or mood have to be inferred. In spite of the advances in signal analysis to date, speech recognition involves an inference and decision process; in other words, the phoneme or word identity in the utterance cannot be directly measured without hypothesis and decision.

Measurement dimensions that are used in the study of acoustic-phonetic properties of a spoken language include such parameters as the short-time energy, the zero-crossing or level-crossing rate, voicing onset, and so on. Probably the most important feature dimension of speech is the short-time spectral envelope, which encapsulates the key characteristics of the articulatory apparatus (e.g., the resonant frequencies of the vocal tract or formants) that caused the realization of the speech sound. Short-time spectral estimation and representation is, thus, considered the core of speech analysis for speech recognition.

Short-time Fourier analysis performed on consecutive blocks of data is the most rudimentary form of short-time spectral analysis for speech. Short-time Fourier analysis, however, produces a result with raw information that may not be most suitable for linguistic inference. Two measures are usually taken. One further smooths the Fourier spectrum along the frequency axis by averaging adjacent frequency components in a weighted manner (e.g., a triangular weight vector) and the other simply resorts to a filter-bank implementation with an embedded nonlinearity and spectral smoothing. We shall denote the short-time spectral estimate of the speech signal by $S(\omega)$, where ω is the normalized frequency. Fig. 10 shows a schematic using a filter bank to obtain the short-time speech spectral envelope. Note that it is now customary to employ a "perceptually motivated" frequency band structure to approximate the human auditory system. "Mel-frequency" or "Bark scale" relates to a human's subjective perception of frequency (in sinusoid), and the notion of critical bandwidth and auditory masking provides a guideline in incorporating the nonuniform spectral resolution in our auditory system into the analysis procedure.

Another major branch of spectral analysis for speech is the autoregressive or all-pole modeling method [31]. Let $s(n)$,

$n = 1, 2, \ldots, N$ be a block of speech data. In autoregressive modeling, $s(n)$ is assumed to be generated by an autoregressive mechanism or source, i.e.,

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \tag{9}$$

where a_k, $k = 1, 2, \ldots, p$ are the autoregressive parameters, also often referred to as the predictor coefficients, p is the order of analysis, and $u(n)$ the innovation sequence or driving function, an independent and identically distributed (i.i.d.) process, with G as the gain parameter affecting the amplitude of the speech signal. Equation (9) can be interpreted as an assumption that each speech sample can be predicted from past samples (as a linear combination), and, thus, this modeling technique is also referred to as linear prediction or linear predictive coding (LPC). Given a sequence of data $\{s(n)\}$, LPC analysis aims at finding a set of parameters $\{a_k\}$ that minimizes the difference between the actual value and the predictive value, i.e., the prediction error

$$e(n) = s(n) - s'(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{10}$$

averaged over the (windowed) data block. Expressed in the z-transform, (10) becomes

$$S(z) = E(z)/A(z) \tag{11}$$

where

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \tag{12}$$

is a pth order polynomial. When the process is indeed an autoregressive process and the prediction error is minimized by a proper choice of $\{a_k\}$, the prediction error will equal $Gu(n)$. The all-pole model power spectrum $1/|A(\omega)|^2$ approaches the speech power spectrum $|S(\omega)|^2$, within a scaling constant (the gain term), when the order p is large (assuming large N), and provides an estimate of the short-time spectral envelope when p is relatively small (usually 10–16 for speech).

As with the filter-bank analysis, perceptual attributes can be incorporated in the linear prediction analysis framework. Examples of such attempts are Mel-scale LPC and perceptual linear prediction (PLP) [32].

Another important attribute in auditory perception is the nonlinear compression of energy, which leads to the consideration of the log power spectrum $\log|S(\omega)|^2$ in most speech-related processing algorithms. The Fourier series representation of $\log|S(\omega)|^2$ can be expressed as

$$\log|S(\omega)|^2 = \sum_{n=-\infty}^{\infty} c_n e^{-jn\omega} \tag{13}$$

where $c_n = c_{-n}$ are real and often called the cepstral coefficients, or simply the cepstrum. Note that c_0 is the average log power spectrum. Depending on the way $S(\omega)$ is estimated, the corresponding cepstrum may possess somewhat different properties. If $S(\omega)$ is modeled as an all-pole spectrum as in the LPC analysis, the corresponding cepstrum is called an LPC cepstrum. When $S(\omega)$ is obtained with a Mel-scaled filter-bank, the cepstrum is then called a Mel-cepstrum. Studies exist to compare these derivatives of the cepstral representation for speech recognition and statistical modeling.

It should be noted that these short-time spectral analysis methods are applied to consecutively windowed speech segments, called "frames," resulting in a sequence of short-time spectral envelope estimates. The sequence with its time-dependent variation, thus, defines the basic speech pattern of a spoken utterance.

1) Distortion Measures and Parameter Representations: Speech analysis produces what can be considered as a raw parametric representation of the feature. Since our goal in speech recognition and understanding is to be able to differentiate one linguistic event from another, discussions of feature representation must involve the measure that we employ to assign dissimilarity between any two observed feature parameter vectors. For spectral parameters, this is usually called the spectral distortion measure or simply the distortion measure.

In order to define a proper distortion measure for speech recognition, several factors have to be taken into account. Ideally, the distortion measure must reflect proportionally the perceptual difference judged by human listeners. For automatic speech recognition, it should also result in high correlation with the linguistic distinction between the two spectral parameter vectors that are being compared. While the ultimate dissimilarity measure that possesses these properties is still not at hand, studies in the past have converged to the use of cepstrum and cepstrum-related distances.

Consider two power spectra, $|S(\omega)|^2$ and $|S'(\omega)|^2$, under comparison. Let

$$V(\omega) = \log|S(\omega)|^2 - \log|S(\omega)|^2. \tag{14}$$

The set of L_p norm defined on $V(\omega)$ is a natural choice in defining the distortion measure between the two power spectra

$$d_p(S, S')^p = \int_{-\pi}^{\pi} |V(\omega)|^p \, d\omega. \tag{15}$$

For $p = 1$, it defines the mean absolute log spectral distortion. For $p = 2$, it defines the rms log spectral distortion, which has become prevalent in many speech-processing algorithms.

By applying Parseval's theorem, we can relate the L_2 cepstral distance to the rms log spectral distortion

$$d_2(S, S')^2 = \sum_{n=-\infty}^{\infty} (c_n - c'_n)^2 \tag{16}$$

where c_n and c'_n are the cepstral coefficients of $|S(\omega)|^2$ and $|S'(\omega)|^2$, respectively. The L_2 norm above usually is truncated to a finite number of terms, say, D, and the finite-dimensional cepstral vector $\mathbf{c}^t = (c_1, c_2, \ldots, c_D)$

is, thus, used as a representation of the feature vector for speech recognition. Several other representations of the short-time speech spectrum have been proposed and studied. The cepstrum, however, remains a prevalent representation in many speech-processing algorithms.

2) Post-Processing of Speech Feature Representations: Measurements or estimated parameters of the speech feature contain inevitably "noise" such as estimation error, model mismatch, and so on. These noisy components appear in the short-time spectral sequence as unreliable fluctuation, which would reduce the effectiveness in pattern matching if not properly treated. Postfiltering on the parameter sequence over time is, thus, advisable in removing some of these "noisy" components. Various techniques such as the Slepian filter and the Legendre polynomial [33], RASTA [34], and similar bandpass filters have been reported to bring about good results.

Another aspect of the short-time spectral feature that can lead to an improved performance is the dynamics of the time-varying parameter sequence. A short-time spectral estimate usually is obtained within a short time window, independent of the adjacent data blocks. However, the rate of change in the parameter sequence (a form of dynamic feature) is believed to have a major perceptual as well as cognitive significance. Additional parameters that are derived from the (static) feature representation and bear the dynamic characteristics of speech have been suggested and shown to help the automatic speech-recognition performance. These dynamic feature representations include the delta-cepstrum (a first-order difference of the short-time cepstral sequence) [35], the delta-delta-cepstrum (a second-order difference), delta-energy (a first-order difference of the short-time energy parameter), delta-delta-energy (a second-order difference), etc. These dynamic feature representations can also be considered results of higher order polynomial data fitting on the short-time (static) feature representation [36]. The parameter window for polynomial data fitting can have a span over several "frames," thus, extending the representation beyond the strictly "short-time frame" range.

C. Acoustic Modeling [40]

Acoustic modeling aims at finding the probabilistic behavior of the given data, expressed in the form of $P(\boldsymbol{X}|\Lambda)$. [Here, we use $P(\boldsymbol{X}|\Lambda)$ in lieu of $P_{\Lambda}(\boldsymbol{X}|C)$ without ambiguity because the class label is implied in the context.] This is often referred to as probability distribution estimation; i.e., finding the parameter Λ in a certain optimal sense to define the distribution $P(\boldsymbol{X}|\Lambda)$. For automatic speech recognition, the first issue to be resolved is the functional form of the distribution that best describes the probabilistic nature of the speech signal. Understanding of the speech variability helps to determine the appropriate form before estimation of the parameter Λ can take place.

1) Probability Distributions for Speech: The statistical method, as discussed in the previous sections, requires that a proper, usually parametric, distribution form for the observations be chosen in order to implement the MAP decision. Using the task of isolated-word speech recognition as an example, we have to determine the distribution form for the speech utterance of each word before we employ an estimation method to find the values of the parameters.

What is the right distribution form for speech utterances? This question involves two essential aspects: finding the speech dimensions that carry the most pertinent linguistic information and deciding how to statistically characterize the information along the chosen dimensions. We discuss these issues in this section.

Speech is a time-varying signal. When we speak, our articulatory apparatus (the lips, jaw, tongue, and velum) modulates the air pressure and flow to produce an audible sequence of sounds. Although the spectral content of any particular sound in speech may include frequencies up to several thousand hertz, our articulatory configuration (the vocal-tract shape, the tongue movement, etc.) often does not undergo dramatic changes more than ten times per second. During the short interval where the articulatory configuration stays somewhat constant, a region of "quasi-stationarity" in the produced speech signal can often be observed. This is the first characteristic of speech that distinguishes it from other random, nonstationary signals. The temporal variation is manifested in several ways: the timing of voicing onsets, the vowel duration, etc. The short-time signal analysis discussed above produces representations of speech in a sequence of parameter vectors containing these characteristics. For speech recognition, however, certain kinds of temporal variation are irrelevant to the linguistic distinction between utterances. For example, most of the variation due to speaking rate changes is not going to alter the linguistic content of the utterance (although it may have semantic and pragmatic implications that are not normally considered part of the speech-recognition task). Representations of speech for recognition purposes, thus, have to take this into account and include the ability to accommodate irrelevant variations or to suppress them.

Furthermore, speech is not a memoryless process due to articulatory and phonotactic constraints. According to the phonological rule of a language, there is a certain dependency between sound pairs that occur in sequence: some occur more often than others, while some are simply nonexistent in the language. The speech model or distribution needs to have provisions to permit characterization of this sequential structure, ideally in a manner consistent with the slowly varying nature (i.e., "quasi-stationarity") of the speech signal.

2) Speech Model: Based on the above characterization of the speech signal, a reasonable speech model or distribution should have the following three components. First, at an interval on the order of 10 ms, short-time measurements are to be made along the pertinent speech dimensions that best carry the relevant information for linguistic distinction. These dimensions determine the observation space in which the distribution is to be defined. This is accomplished in signal analysis and the choice of representation (Section III-B). Second, the existence of the quasi-stationary region suggests that the neighboring

short-time measurements on the order of 100 ms need to be simultaneously considered, either as a group of independently and identically distributed observations or as a segment of a (perhaps nonstationary) random process covering two quasi-stationary regions. Third, a mechanism that describes the sound change behavior among the sound segments in the utterance is needed. This characterization takes into account the implicit structure of the utterance (words, syntax, and so on) in a probability distribution sense.

3) Hidden Markov Model (HMM): An HMM provides a simple means to characterize speech signals according to the above discussion [37]. Consider a first-order N-state Markov chain governed by a state transition probability matrix $A = [a_{ij}]$, where a_{ij} is the probability of making a transition from state i to state j. Assume that at $t = 0$ the state of the system q_0 is specified by an initial state probability $\pi_i = P(q_0 = i)$. Then, for any state sequence $\boldsymbol{q} = (q_0, q_1, \ldots, q_T)$, the probability of $\boldsymbol{q}$ being generated by the Markov chain is

$$P(\boldsymbol{q}|A, \pi) = \pi_{q_0} a_{q_0 q_1} a_{q_1 q_2} \cdots a_{q_{T-1} q_T}. \tag{17}$$

Suppose the system, when at state q_t, puts out an observation $\boldsymbol{x}_t$ according to a probability density function $b_{q_t}(\boldsymbol{x}_t) = p(\boldsymbol{x}_t|q_t)$, $q_t = 1, 2, \ldots, N$. The HMM used as a distribution for the speech utterance $\boldsymbol{X}$ is then defined as

$$\begin{aligned} P\left(\boldsymbol{X}|\pi, A, \{b_j\}_{j=1}^N\right) &= P(\boldsymbol{X}|\Lambda) = \sum_{\boldsymbol{q}} P(\boldsymbol{X}, \boldsymbol{q}|\Lambda) \\ &= \sum_{\boldsymbol{q}} P(\boldsymbol{X}|\boldsymbol{q}, \Lambda) P(\boldsymbol{q}|\Lambda) \\ &= \sum_{\boldsymbol{q}} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1} q_t} b_{q_t}(\boldsymbol{x}_t) \end{aligned} \tag{18}$$

where $\Lambda = (\pi, A, \{b_j\}_{j=1}^N)$ is the parameter set for the model. Fig. 11 depicts the concept of an HMM, as a measure for probabilistic functions of a Markov chain. As shown in the illustration, each state is associated with a random process, governed by a distribution.

As can be seen in (18), $\{b_{q_t}\}$ defines the distribution for short-time observations and A characterizes the behavior and interrelationship between various states of the speech generation process. In other words, the structure of an HMM provides a reasonable means for characterizing the distribution of a speech signal. Normally N, the total number of states, is much smaller than T, the time duration of the speech utterance. The state sequence $\boldsymbol{q}$ displays a certain degree of stability among adjacent q_ts due to the above-mentioned "quasi-stationarity." The use of HMM as speech distributions has been shown to be practically effective.

The form of the in-state observation density $b_{q_t}(\boldsymbol{x}_t)$ needs to be specified. Different choices of speech dimensions for the observation space may require different forms of the in-state observation distribution. One general form, namely the mixture Gaussian density [20], is commonly employed

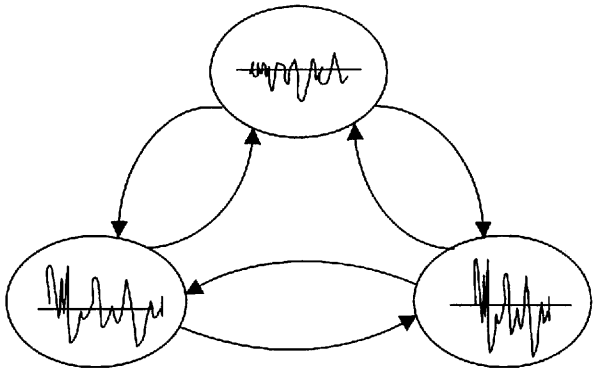


Fig. 11. Illustration of a hidden Markov model (HMM).

due to its ability to approximate arbitrary density functions. A mixture density has the form

$$b(\boldsymbol{x}) = \sum_{j=1}^{K} \gamma_j f_j(\boldsymbol{x}) \tag{19}$$

where $f_j(\bullet)$ is the jth component kernel function, usually a Gaussian density, and γ_j is the weight of the density component. (Note that we have dropped the state index in the above generic expression for a mixture density without ambiguity.) By increasing the number of mixture components, a mixture Gaussian density can approximate any density function with arbitrary precision. The tradeoff is obviously among the closeness in approximation, the increase in number of parameters, and the related parameter estimation reliability, which is a function of the amount of training data. In any event, there is always a possible discrepancy between the estimated distribution and the true data distribution. This notion is important in the following discussion of discriminative method (Section III-C5).

4) Model Parameter Estimation: Once the speech model form is chosen, the parameter set Λ that defines the model is to be estimated from a given set of data. This process is often referred to as training in the context of automatic speech or pattern recognition. For the simple case of isolated word or discrete utterance recognition in which an utterance constitutes an observation of a class, each token in the training set carries a label of the class identity. Normally, a class-dependent model or distribution is estimated from the data of the same class according to some well-known statistical estimation criteria, "maximum likelihood" (ML) being one of the most prevalent ones. Let $\Omega_X = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_L\}$ be the training set for a particular class of observations. Maximum-likelihood estimation is to find the parameter set Λ_{ML} such that

$$\Lambda_{\mathrm{ML}} = \arg\max_{\Lambda} P(\Omega_X|\Lambda) \tag{20}$$

where

$$P(\Omega_X|\Lambda) = \prod_i P(\boldsymbol{X}_i|\Lambda)$$

assuming that the utterances in the training set are independent. The estimated model parameter set Λ is then associated

with each individual word class. For an M-word vocabulary, M such parameter sets are to be estimated for use in the recognizer.

The Baum–Welch algorithm [38], [20] accomplishes likelihood maximization in a two-step procedure, known as "reestimation." Based on an existing model Λ (or a properly initialized model), the first step of the algorithm transforms the objective function $P(\Omega_X|\Lambda)$ into a new function $Q(\Lambda', \Lambda)$ that essentially measures a divergence between the initial model Λ' and an updated model Λ. The Q function is defined, for the simplest case, as

$$Q(\Lambda, \Lambda') = \sum_{\boldsymbol{q}} P(\Omega_X, \boldsymbol{q}|\Lambda')\log P(\Omega_X, \boldsymbol{q}|\Lambda) \tag{21}$$

where $P(\Omega_X, \boldsymbol{q}|\Lambda)$ can be derived according to (18) and (20). It can be shown that $Q(\Lambda', \Lambda) \geq Q(\Lambda', \Lambda')$ implies $P(\Omega_X|\Lambda) \geq P(\Omega_X|\Lambda')$. Therefore, the second step of the algorithm involves maximizing $Q(\Lambda', \Lambda)$ as a function of Λ to obtain a higher, improved likelihood. These two steps iterate interleavingly until the likelihood reaches a fixed point.

The ML method is, however, not the only possible choice for solving the estimation problem. An in-depth discussion of various estimation criteria can be found in [37]. It should be pointed out that the ML method does not usually lead to a minimum error rate performance for the recognizer. As discussed above, this is due to 1) the likely mismatch between the chosen distribution form (HMM in the present case) and the actual speech data and 2) the finite training (known) data set, which is often inadequate.

5) Discriminative Training: As discussed earlier, classifier design by distribution estimation often does not lead to an optimal performance. The problem is that in most situations, the estimated probabilities deviate from the true probabilities and the exact MAP rule cannot be implemented. In addition, when the assumed form of the distribution is different from the true one, the optimality of the estimated distribution has little to do with the optimality of the classifier, particularly in terms of recognition error rate. An attempt that has developed over the past few years to overcome the fundamental limitations of the traditional approach based on distribution estimation is to directly formulate the classifier design problem as that of classification error rate minimization. This approach is called "discriminative training," in which the goal of training is to be able to correctly discriminate the observations for best recognition/classification results rather than to fit the distributions to the data [39].

Consider a set of *discriminant functions* $g_i(\boldsymbol{X}; \Lambda)$, $i = 1, 2, \ldots, M$ defined by the parameter set Λ. In its simplest form for our present discussion of the HMM technique, $g_i(\boldsymbol{X}; \Lambda)$ can take essentially the same form as (18), i.e.,

$$\begin{aligned} g_i(\boldsymbol{X}; \Lambda) &= P\left(\boldsymbol{X}|\lambda^{(i)}\right) \\ &= P\left(\boldsymbol{X}\middle|\pi^{(i)}, A^{(i)}, \left\{b_j^{(i)}\right\}_{j=1}^{N}\right) \end{aligned} \tag{22}$$

where the superscript i denotes the parameter set identity associated with word (class) i in the vocabulary. It is important to note that in this formulation, we no longer consider only the estimation of distribution for a class of observations, but take into account the entire parameter set of the classifier Λ, $\Lambda = \{\lambda^{(i)}, i = 1, 2, \ldots, M\}$ in the optimization process. The discriminant $g_i(\boldsymbol{X}; \Lambda)$ can be any reasonable functions. (Discussion of its optimality is beyond the scope of this paper.) The choice of HMM of (18) is a reasonable one (perhaps the best we have so far), as discussed previously. The classifier/recognizer is operating under the following *decision rule*:

$$C(\boldsymbol{X}) = C_i \qquad \text{if } g_i(\boldsymbol{X}; \Lambda) = \max_j g_j(\boldsymbol{X}; \Lambda). \tag{23}$$

The goal of classifier design is again to achieve the minimum error probability based on the loss function defined in (4).

The difficulty associated with the discriminative training approach lies in the derivation of an objective function that has to be consistent with the performance measure (i.e., the error rate) and also suitable for optimization. The error rate based on a finite data set is a piecewise constant function of the classifier parameter Λ and, thus, a poor candidate for optimization by a simple numerical search method. An *embedded smoothing* for a loss function that is a reasonable approximation to the error probability has been proposed [39].

a) Optimization criterion: The smoothed optimization criterion is a function of the *class discriminant functions* $g_i(\boldsymbol{X}; \Lambda)$, $i = 1, 2, \ldots, M$. We assume that the discriminant functions are nonnegative. The key to the new error criterion is to express the operational decision rule of (23) in a functional form. There exist in this regard many possibilities, one of which is a misclassification measure taking the following form:

$$d_i(\boldsymbol{X}) = -g_i(\boldsymbol{X}, \Lambda) + \left\{\left(\frac{1}{M-1}\right)\sum_{j, j\neq i}[g_j(X, \Lambda)]^{\eta}\right\}^{1/\eta} \tag{24}$$

where η is a positive number. This misclassification measure is a continuous function of the classifier parameters Λ and attempts to enumerate the decision rule. For an ith class utterance $\boldsymbol{X}$, $d_i(\boldsymbol{X}) > 0$ implies misclassification and $d_i(\boldsymbol{X}) \leq 0$ means correct decision. When η approaches ∞, the term in the bracket becomes $\max_{j, j\neq i} g_j(\boldsymbol{X}, \Lambda)$. By varying the value of η, one can take all the competing classes into consideration, according to the individual numeric significance, when searching for the classifier parameter Λ.

To complete the definition of the objective criterion, the misclassification measure of (24) is embedded in a smoothed zero–one function, for which any member of the sigmoid function family is an obvious candidate. A general form of the loss function can then be defined as

$$L_i(\boldsymbol{X}; \Lambda) = \ell_i(d_i(\boldsymbol{X})) \tag{25}$$

where ℓ is a sigmoid function, one example of which is

$$\ell_i(d) = \frac{1}{1 + \exp(-\rho d + \theta)} \tag{26}$$

with θ normally set to zero. Clearly, when $d_i(\boldsymbol{X})$ is much smaller than zero, which implies correct classification, virtu-

ally no loss is incurred. When $d_i(\boldsymbol{X})$ is positive, it leads to a penalty that becomes essentially a classification/recognition error count. Finally, for any unknown $\boldsymbol{X}$, the classifier performance is measured by

$$L(\boldsymbol{X};\ \Lambda) = \sum_i L_i(\boldsymbol{X};\ \Lambda)\mathbf{1}(\boldsymbol{X} \in C_i) \qquad (27)$$

where $\mathbf{1}(\bullet)$ is the indicator function.

This three-step definition emulates the classification operation as well as the performance evaluation in a smooth functional form, suitable for classifier parameter optimization. Based on the criterion of (27), we can choose to minimize one of two quantities for the classifier parameter search: one is the expected loss and the other the empirical loss. The method of generalized probabilistic descent (GPD) has been proposed to achieve this goal. Application of the direct error minimization method to HMMs for speech recognition can be found in [40].

D. Acoustic Modeling for Large-Vocabulary Continuous Speech Recognition

The Bayes decision theory applies straightforwardly to the simple case of isolated word recognition, as discussed above. For large-vocabulary continuous speech recognition, however, several issues arise, mainly due to the increased variability, in all stages of the speech production chain and the complexity in the recognition process.

In a continuous speech utterance, there exists an abundance of the so-called coarticulation phenomenon. When a sequence of sounds is being uttered, our articulatory apparatus normally does not produce each sound individually. It adjusts its configuration in anticipation of the following sounds, often causing partial articulation or substantial variation to the sounds. In many cases, the colloquial form of a complete sentence may be quite different from the phoneme sequence prescribed by the lexicon. (For example, try to speak "Did you eat yet?" and hear how it differs from the dictionary.) Increase in variability means the need of a more complex system, using speech models with many more parameters in order to maintain the needed modeling performance.

The complexity of a recognition system is essentially proportional to the number of classes to be recognized. For continuous speech recognition, the number of classes can vary, depending on the choice of "unit." If "word" is chosen as the fundamental building block of the spoken language involving, say, a 20 000 word vocabulary, the necessary number of word models is then at least 20 000. And if the above-mentioned variability increase is taken into account, the number of word models can easily exceed this figure. This is obviously a difficult task, even in training. To train 20 000 good word models, the number of training tokens will be in the millions. Collecting data itself becomes a hard task to accomplish. Other alternatives of the recognition unit for acoustic modeling in large-vocabulary speech recognition include syllables (on the order of tens of thousands in number), phonemes (~50), demisyllables (syllable doublet, ~2000), and acoustic units. (Acoustic units are "typical" speech sounds automatically clustered from the speech data, with a varying number in the set according to the needed resolution in sound differentiation.)

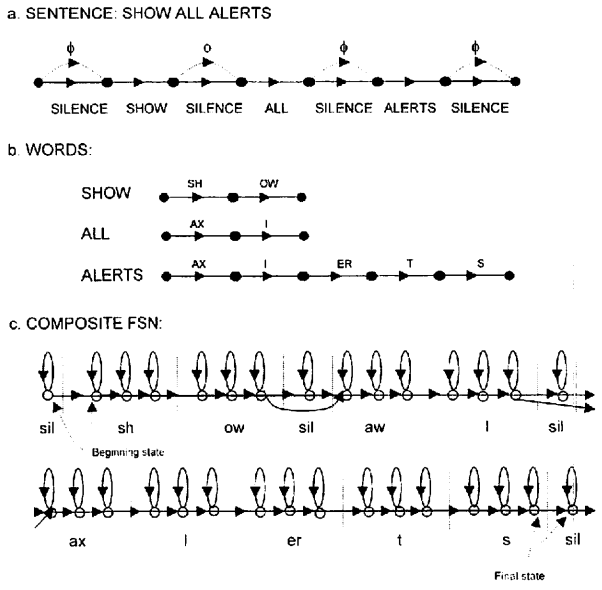


Fig. 12. A composite finite state network representation for sentence "Show all alerts" (after [37]).

One popular choice of model unit is the set of phoneme-like units, which can be considered phonemes with acoustic variation due to various contexts. For example, the vowel /a/ can be represented by a number of acoustic models, qualified by the context it appears in, such as silence-/**a**/-/b/, /k/-/**a**/-/t/, and so on as its variants. In this manner, many more context-dependent models are enlisted. Although only 43 or so phonemes are customarily included in American English, an automatic speech recognition system usually uses several thousands of these phoneme-like units.

By way of illustration, Fig. 12 shows the representation and creation of a sentence "show all alerts" in phoneme-like units expressed in a finite state network. Fig. 12(a) depicts the word sequence with possible insertions of pauses between words in a network. Fig. 12(b) shows the lexical representation of each of the words in the sentence. Fig. 12(c) displays the model for silence (a single-state HMM) and a composite finite state network connecting all the words and silence together in terms of the states in the phoneme-like acoustic models. When context-dependent units are used, each phoneme-like model in the representation will need to have the context designation.

These models can be automatically trained using the same Baum–Welch algorithm or the segmental k-means algorithm [37]. With the representation of a composite finite state model, the correspondence between the models (parameters of which are to be estimated) and the speech utterance becomes explicit much the same way as in the isolated-word case. The same training formulation, thus, applies.

The tradeoff between high acoustic resolution and low estimation reliability is an issue that needs particular care in large-vocabulary speech modeling. Recent research in

acoustic modeling for large vocabulary speech recognition has pushed the context dependency to the level of "quinphone" (a phoneme-like unit with dependency on the two preceding phonemes and two subsequent phonemes). With the high number of potential quinphone units to include, unit selection, thus, becomes critical to avoid loss of reliability in the estimated results, causing offset to the gain from the increased acoustic resolution. Differential likelihood (or the Kullback–Liebler distance for HMM) [41], [42] is often used for this purpose. Another technique to gain reliability in parameter estimate is the use of tying. Tying forces models of different units to share part of the network topology and the parameters, thereby reducing the total number of model parameters.

E. Language Modeling—From Pronunciation to Grammar

When the speech signal is converted into a (discrete) sequence of (unitary class) symbols, it has to conform to the constraints imposed by the relationship among these symbols. In language, some words follow a particular word more likely than others (i.e., the existence of a syntactical or grammatical relationship among words in expressions), and according to the dictionary, phonemes do not arbitrarily follow each other (i.e., due to phonological rules and the lexicon). A spoken language system needs to invoke the knowledge of pronunciation, lexicon, and syntax in order to be able to perform satisfactorily in decoding the speech signal into a text sequence ready for further interpretation. Models that encapsulate these aspects of knowledge are indispensable. We discuss the prevalent considerations and structures in this section.

1) Model Structure: In the current probabilistic framework, the language that governs the outcome of a linguistic event is modeled as a discrete density $P(W)$, where $W = (w_1, w_2, \ldots, w_L)$ is a word sequence. The density function $P(W)$ assigns a probability to a particular word sequence W, depending on how likely it is to appear in the task. A sentence with words appearing in a grammatically correct manner is more likely to be spoken than a sentence with an ungrammatical structure, and, therefore, is assigned a higher probability. The language structure, rules, and convention are, thus, integrally considered in the probability assignment. Statistical language modeling is to estimate $P(W)$ from a given set of sentences, or corpus.

Whether or not a language can be adequately described in a statistical model has been subject of debate for some time. Traditional linguistics considers language to comprise a set of rules that define the proper expression in terms of syntax, semantics, and even pragmatics. It is this set of commonly accepted and recognized rules that facilitate communication (which by itself means sharing of information). The rules are fundamentally deterministic. Parsing is to map a word sequence into a structure that can be interpreted according to the rules for further understanding. The complication comes, however, from the interaction between the structure (i.e., a sentence that can be properly parsed) and the message (words that generate the intended meaning). This interaction creates not only variations in expression, but also potential confusion and difficulty in parsing. Can a parser properly parse a sentence containing significant amounts of word and syntax errors for comprehension?

Similarly, can proper lexical modeling resolve ambiguities in pronunciation, which often displays wide variation in actual acoustic realizations of words and phrases? What is the right model structure or, more precisely, the right mathematical representation that allows people to characterize the variation? These questions point to an ensemble of problems that are still open—representation of various linguistic structures from message to articulation.

While research is still active in the pursuit of structural representations, a number of propositions exist [43], among them context-free grammar (CFG), finite state grammar (FSG), probabilistic CFG, probabilistic FSG, and so on. The finite state grammar, due to its computational synergy with the HMM, is widely used in many engineering realizations of spoken language systems.

2) Probabilistic Finite State Language Model: A finite state grammar is specified by the vocabulary V, the current observation $w_t \in V$, the current state $G_t = (w_{t-N}, w_{t-N+1}, \ldots, w_{t-1})$, and the next state function $G_{t+1} = \phi(G_t, w_t)$. The next state function defines the evolution of state given the observations. One simple form of the next state function is

$$\begin{aligned} G_{t+1} &= \phi(G_t, w_t) \\ &= (w_{t-N+1}, w_{t-N+2}, \ldots, w_t) \end{aligned} \tag{28}$$

which is just a shift of the word sequence window of size N. A general and yet convenient way to visualize an FSG is the word network. Recall an example of the grammar network, shown in Fig. 2, specifying many possible ways to express the notion of trying to get information about a flight. The next state function can be probabilistic. The word window size N is called the order of the finite state model, which invokes a syntactic rule or assigns a measure of probability to a word given its state specification. In other words, a finite state language model defines $P(w_t|G_t)$.

The simplest form of a first-order language model is the word-pair grammar in which $P(w_t|G_t) = P(w_t|w_{t-1}) = 1$, if the syntactic rule permits w_t to follow w_{t-1}; otherwise, $P(w_t|w_{t-1}) = 0$. Similarly, one can define a triplet language model, and so on.

In speech recognition and understanding, a probabilistic N-gram FSG has found widespread use due to its implementational ease and consistency with the structure of an HMM. An N-gram language model is a statistical model that assigns a probability measure to a word sequence using the following approximation:

$$P_N(W) = \prod_{t=1}^{T_e} P(w_t|w_{t-1}, w_{t-2}, \ldots, w_{t-N+1}). \tag{29}$$

The conditional probability $P(w_t|w_{t-1}, w_{t-2}, \ldots, w_{t-N+1})$ can be estimated by the simple relative frequency approach—counting the frequency of occurrences of the

partial word sequences. To circumvent the problem of small probability estimation (e.g., unseen events in a limited training set), a backoff strategy based on the Good–Turing estimator is often incorporated [12].

Associated with a statistical language model is the entropy or perplexity that measures the complexity of a language that the language model is designed to represent [42]. Such a measurement in theory has to be made over all expressions of the language. In practice, the entropy of a language according to a language model $P_N(W)$ is measured via a set of sentences and is defined as

$$H = \sum_{W \in \Omega} P_N(W) \log P_N(W) \tag{30}$$

where Ω is a set of sentences of the language. The perplexity, which is also called the average word-branching factor, is defined as

$$B = 2^H. \tag{31}$$

The perplexity, even though it is an estimated quantity [such as the test-set perplexity obtained via (30)], provides an indication of the general complexity and, thus, the difficulty, of a given language task.

It is possible to extend the fixed-order FSG to a tree-structured variable-order FSG. A variable-order FSG in general achieves lower estimated perplexity than a fixed-order language model. However, when applied to speech recognition and understanding, the performance difference in error rate is usually insignificant. (See discussions below.)

F. Pattern Matching and Search

Pattern matching refers to the computational procedure to evaluate the similarity between the patterns that are being compared. Such a step is necessary in every pattern-recognition system. The basis of pattern matching is the underlining distortion measure that the system designer chooses *a priori*. Considerations in choosing a distortion measure have been discussed in Section III-B1. As we have pointed out, a distortion measure is often associated with some notion in statistical distribution, and the amount of distortion can be interpreted as the negative log likelihood. In speech recognition, the procedure of pattern matching, thus, can be considered either as the computation of distortion/dissimilarity between two patterns or the computation of likelihood of a source (represented by a model) that produced the observed speech pattern, followed by a search and decision process.

Two basic techniques are available for speech pattern matching. One is called the Viterbi algorithm and the other the forward–backward algorithm. These two techniques correspond to the computation of the two essential quantities, respectively, in the formulation of a statistical model for speech. The Viterbi algorithm allows efficient computation of $\max_q P(\boldsymbol{X}, \boldsymbol{q}|\Lambda)$ and the forward–backward algorithm leads to linear-time computation of $P(\boldsymbol{X}|\Lambda)$.

1) Dynamic Programming and the Viterbi Algorithm: Evaluating the dissimilarity between two speech patterns is more sophisticated than most of the traditional pattern-matching technique that operates in a simple fixed-dimensional vector space. One main factor that contributes to this added sophistication is that speech pattern is a temporal sequence that may be compressed or stretched in time due to speaking rate changes. Such a speaking rate variation would usually result in a nonuniform duration in the phonemic elements of the utterance during articulation. To obtain a meaningful pattern-matching score between two speech patterns of unequal length, some form of normalization is necessary. This gives rise to the use of dynamic time warping in defining a dissimilarity measure, beyond the distance calculation between two short-time vector representations of speech.

Consider two speech patterns: $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{T_x})$ and $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{T_y})$. Let $d(\boldsymbol{x}, \boldsymbol{y})$ be the distortion measure between two short-time representations of speech $\boldsymbol{x}$ and $\boldsymbol{y}$. We shall also use the following simplified expression $d(\boldsymbol{x}_i, \boldsymbol{y}_j) = d(i, j)$. Also, let $i_x = \phi_x(k)$ and $i_y = \phi_y(k)$ be two time-warping functions that relate the time scale of $\boldsymbol{X}$, i_x, and that of $\boldsymbol{Y}$, i_y, respectively, to an independent time index k. Given the two time-warping functions, one way to define the dissimilarity measure between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is

$$d_\phi(\boldsymbol{X}, \boldsymbol{Y}) = \min_{\phi} \sum_{k=1}^{T} d(\phi_x(k), \phi_y(k)). \tag{32}$$

Obviously, the warping function has to satisfy certain constraints to prevent meaningless match; for example, a close match between "we" and "you" would result if ϕ is not constrained to be monotonically nondecreasing.

To find the minimum in (32), proposals by Vintsyuk [9] and by Sakoe and Chiba [10] in using dynamic programming techniques prove very effective. The principle of dynamic programming is that an overall (global) optimal path must also be locally optimal. The principle leads to a procedure to increment an accumulative optimal path up to time t by advancing to the point that would be optimal at time $t + 1$. It translates a search for global optimality into a search for local optimality for the above linear optimization problem. Dynamic programming, thus, achieves the optimization objective in linear, rather than exponential, time.

The formulation in (32) has direct correspondence in solving the evaluation problem in hidden Markov modeling. Note that the state sequence $\boldsymbol{q}$ in the joint state-observation likelihood function $P(\boldsymbol{X}, \boldsymbol{q}|\Lambda)$ plays the same role as the time-warping function ϕ, and the optimization objective $\max_q P(\boldsymbol{X}, \boldsymbol{q}|\Lambda)$ is identical to that in (32) (recall that the distortion can be viewed as negative log likelihood). The reference pattern is now the HMM whose state distributions form essentially a "pattern," albeit with explicit probability assignments. The same dynamic programming algorithm is obviously directly applicable in the calculation of $\max_q P(\boldsymbol{X}, \boldsymbol{q}|\Lambda)$. In this context, which is considered maximum-likelihood "decoding" of the code sequence $\boldsymbol{q}$ as in data communication, it is customarily called the Viterbi algorithm.

2) The Forward–Backward Algorithm: One of the fundamental problems in hidden Markov modeling is to evaluate

$P(\boldsymbol{X}|\Lambda)$ of (18) as the likelihood of Λ based on the observation $\boldsymbol{X}$. The difficulty of this evaluation problem comes from the need to sum the state likelihood $P(\boldsymbol{X}, \boldsymbol{q}|\Lambda)$ over all possible state sequences $\boldsymbol{q}$. For an N-state HMM and an utterance sequence of T vectors, the total number of state sequences amounts to N^T, which can be prohibitively high when T is large. A method called the forward–backward algorithm linearizes the exponential computational complexity by making use of the forward and the backward probabilities.

The forward probability $\alpha_t(i)$ is defined as the probability of the partial observation sequence $\boldsymbol{x}_1\boldsymbol{x}_2 \ldots \boldsymbol{x}_t$ (up to time t) and state i at time t, given the model; i.e.,

$$\alpha_t(i) = P(\boldsymbol{x}_1\boldsymbol{x}_2 \ldots \boldsymbol{x}_t, q_t = i|\Lambda).$$

Evaluation of $\alpha_t(i)$ can be accomplished inductively as follows:

1) initialization

$$\alpha_1(i) = \pi_i b_i(\boldsymbol{x}_1), \qquad 1 \le i \le N$$

2) induction

$$\alpha_{t+1}(i) = \left[\sum_{j=1}^{N} \alpha_t(j)a_{ji}\right] b_i(\boldsymbol{x}_{t+1}),$$
$$1 \le i \le N \text{and } 1 \le t \le T-1$$

3) termination

$$P(\boldsymbol{X}|\Lambda) = \sum_{i=1}^{N} \alpha_T(i).$$

This induction procedure requires only N^2T calculations, a linear function in T.

The backward probability can be defined in a similar way

$$\beta_t(i) = P(\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+2} \ldots \boldsymbol{x}_T|\, q_t = i, \Lambda).$$

By using these forward and backward probabilities, many quantities needed in hidden Markov modeling and training can be easily computed. This is the so-called forward–backward algorithm.

3) Search Algorithms for Large-Vocabulary Speech Recognition: Evaluation of the likelihood for the unit models is only the first step in continuous speech decoding. As discussed previously, for large-vocabulary continuous speech recognition, composite models comprising sequences of unitary models are used for "pattern matching." During decoding, the system needs to hypothesize a number of such composite models for likelihood evaluation and recognition decision. Since the number of unitary models can be huge, a naive enlisting based on exhaustive combinatorics will produce prohibitively many hypothesized sequences. Even with the help of a language model to prescreen the sentence hypotheses, the list can still be large and yet has the potential pitfall of excluding the correct ones. A search strategy that combines all levels of likelihood scores and uses them to guide the hypothesis pruning process is, therefore, critical in any practical implementation of a large-vocabulary speech recognition system.

Search algorithms can be categorized into two essential types: best first and breadth first. An algorithm called beam search has been extensively used with good results. It searches the hypothesis space only around the (instantaneous) best path to reduce the amount of likelihood computation. Since the instantaneous best path may not turn out to be the overall best, other strategies may be incorporated to reduce the potential search error. The paper by Ney [44] in this issue provides a thorough discussion of search algorithms.

Another useful search algorithm is the N-best algorithm, which produces not just the best but the top N candidates. The tree-trellis algorithm is particularly noteworthy for its efficiency. With a list of the top N hypotheses, one can invoke higher level or independent knowledge (e.g., a higher order language model or a semantic model) to execute a "multi-pass" strategy to improve the search result. For example, in a task as simple as connected-digit recognition of a credit-card number sequence, after the N top digit sequence hypotheses are produced, one can invoke the error-protection information (e.g., the check-sum digit) embedded in the sequence in making the final recognition decision. The N-best search helps make the use of various levels of knowledge more manageable and efficient, as an integrated "one-pass" search incorporating all the model knowledge at once would be too complex to be realized.

G. Understanding and Dialogue

Having decoded the speech signal into a sequence of words, or a hypothesized sequence of words, a traditional speech understanding system employs a sentence parser to cast the word sequence into a structure to allow syntax verification and inference of meaning. The coupling between parsing and understanding is, however, not a particularly tight one because most parsing algorithms focus on the linguistic structure first, rather than understanding.

At the present time, the goal of automatic understanding of speech is limited to determining the required action based on the speech input. Telephone call routing, which connects a call to a proper destination based on the spoken query, is one such example. Sample sentences collected from a field trial at a banking institution include: "I would like to make a deposit"—to request a connection to the deposit department; "I'd like to borrow money to buy a car"—the loan department; and "My bill does not look right, can I talk to someone?"—the billing department; and so on. Most of these queries involve a single action to be taken by the machine. Another simple speech understanding task that has been attempted is DARPA's Air Travel Information System (ATIS) [45]. In the system, the user talks to the machine to obtain flight information such as "I would like to leave New York for San Francisco on November first, please list the available flights"; "How much does the flight cost from Dallas to Detroit?" In this task, the action to be taken involves the need to cope with the language structure in order to decide which information is to be provided to the user.

For the call routing type of applications, the problem is essentially that of pattern recognition. The observation is the query sentence, which contains a sequence of words. The classes for recognition are the actions (e.g., routing the call to a proper department). There can be several layers of approaches to this problem, depending on the depth of the linguistic inference that the system is designed to pursue. The simplest approach is to assume that in most query sentences, the intended action is going to be expressed in specific terms, spoken in isolation or possibly embedded in a natural utterance. For example, most people probably would instruct the telephone operator/receptionist to connect to a loan department by saying "loan department, please." With the assumption that actions are likely to be expressed in keywords, the system can just employ keyword-spotting techniques to perform the task. This kind of systems is simple to implement. It, however, requires the inclusion of an operator (human) backup when none of the prescribed keywords appears in the utterance or when the system fails to detect the keywords. Another more complex approach that has been attempted takes into account all the words in the utterance, but without paying particular attention to the sequential order of the words. The method of information network [46] or latent semantic analysis [47] has been proposed with reasonable success. These methods use a correlation matrix or network between the actions and the occurrence of words to facilitate the decision process. Compared to keyword-spotting, these methods do not separate *a priori* words that are keywords and those that are not. They implicitly associate a (continuously valued) significance level between the appearance of a word and the intended action. With the added level of complexity, it copes better with the natural utterance input when the number of actions is more than what a system's menu prompt can practically cover or a normal user can remember. The most complex systems for this kind of application may involve use of a parser and attempt to address the semantic aspect of the query. However, it is often more straightforward to resort to a simple mixed-initiative dialogue for query clarification and inference than to attempt to parse the utterance for a thorough understanding.

For voice-enabled applications in information services such as the ATIS task, the system needs to go beyond a simple keyword-spotting scheme. This is due to the complexity of the information to be presented. In many cases, all the query information needed for a proper decision on what flight information to provide is not given in a single utterance. A proper response to a query like "Is there anything cheaper?" requires the knowledge of the history and the reference point in the transaction. It is, therefore, necessary to keep track of the state of dialogue. For applications like ATIS that have a rather constrained set of notions (e.g., departure city, destination, time of departure, time of arrival, date, fare, and so on), the dialogue state can be conveniently represented by a template consisting of the required query field. The interaction between the user and the machine is to first get the template filled with proper information before the end result is presented. The system attempts to extract query information from the sentence. A city name following preposition "to" can be treated as the destination, for example. Any blank in the template or ambiguity in the input would induce a question from the system for further interaction. Clarification and follow-up questions can be initiated based on the information in the template as well as a temporary cache that records intermediate answers.

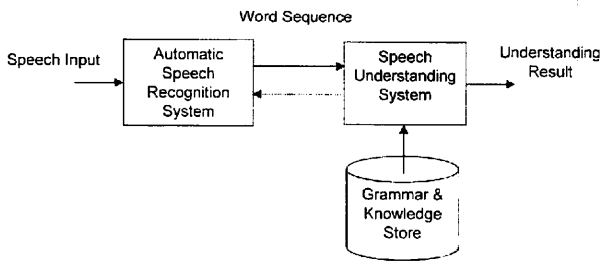


Fig. 13. Block diagram of an understanding system based on speech to word conversion.

Another area of research is also noteworthy. With today's spoken language technology, many applications are being developed, mostly in a trial form to collect the field data for performance improvement and to refine the flow of human–machine interactions in order for the application to deliver a user-friendly experience. The application development effort, nevertheless, can be at times extremely costly, as it often requires manual adjustments in the system design. Software tools that help abate the costly dialogue application development are probably as important as the system technology itself, given the current state-of-the-art of the spoken language technology.

H. A Detection-Based Paradigm for Speech Understanding

In our discussion of speech understanding and dialogue beyond a simple keyword-spotting scheme, the operating assumption is that there is an automatic speech recognizer that is doing its best job in converting the speech signal into a sequence of word, ready to be "understood." Fig. 13 depicts such an architecture. This is an obviously idealized situation, without taking into account the amount of recognition errors in the decoded word sequence. How would the recognition error affect the semantic latency machine or the information theoretic network? And even more interestingly, is there any possibility in giving feedback from the "understanding module" to the speech recognition module such that decoding hypotheses can be properly adjusted and, hopefully, "converge" to the "most correct" word sequence as well as the "most correct" understanding of the utterance? Before a fully integrated decoding and understanding system (with understanding feedback) can be established practically, an alternative paradigm is to use a detection-based approach.

Fig. 14 depicts the fundamental architecture of a detection-based system, which can be applied to phoneme/word recognition as well as linguistic event detection. Each detector aims at detecting the presence of a prescribed event. It can be a phoneme, a word, a phrase, or a linguistic notion such as an expression of date. The detector uses a model for the event (i.e., Λ_0 for the H_0 hypothesis) and an antimodel

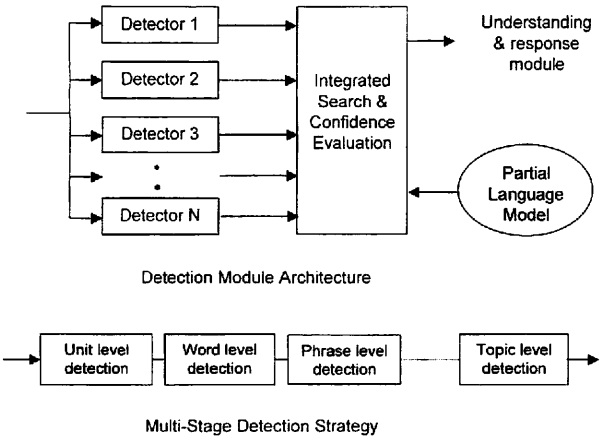


Fig. 14. Architecture of a detection-based speech understanding system.

that provides contrast to the event (i.e., Λ_1 for the H_1 hypothesis). It follows the Neymann–Pearson lemma in that the likelihood ratio is used as the test statistic against a threshold.

The prescribed event is present or detected iff

$$\frac{P(\boldsymbol{X}|\Lambda_0)}{P(\boldsymbol{X}|\Lambda_1)} \geq \sigma. \tag{33}$$

Several issues need to be addressed in this new formulation. First, training of the models and antimodels can follow the usual maximum-likelihood methods, using the properly segmented tokens for Λ_0 and the rest of the observations (sometimes called the garbage) for Λ_1. It can also follow the idea of discriminative training, using the verification error as the optimization criterion. The type I verification error (often called a "miss," or failure to detect the presence of a legitimate event) for a known event $\boldsymbol{X}$ is, with the above verification strategy

$$\ell(d) = \ell\{-\log P(\boldsymbol{X}|\Lambda_0) + \log P(\boldsymbol{X}|\Lambda_1) + \log\sigma\} \tag{34}$$

where ℓ is the sigmoid function as defined previously. The type II error (also called "false alarm," erroneous detection of a spurious event as a legitimate one) can be similarly defined by reversing the signs in the expression of (34), since this is equivalent to a two-class problem. Model parameters can be optimized using as the criterion a combination of the type I and the type II error according to a prescribed operating strategy (i.e., the relative significance of type I and type II errors).

The second issue in this approach is the choice of detection units. Mathematically, the longer the test observation is, the more reliable the detection result will be. The tradeoff depends on the task and the range of variability in various levels of expressions. A reasonable choice of units to be included in the detector set is the words in the vocabulary, key phrases in the task, or important language expressions such as the date and the time, etc. It can also be extended to higher level notions such as a topic, an action, and so on, provided it has an operating expression.

The third issue in this approach is the inclusion of event context, which can help raise the performance of the system in the integrated search after the detectors propose individual decisions. Integration of language models and high-level constraints in the search beyond the detection units is not straightforward and is an active area of research.

Several simple implementations of this new paradigm have shown promises in dealing with natural utterances containing many out-of-vocabulary words (including extraneous sounds such as uh's and um's, partial words, repairs) or out-of-grammar sentences [48]. Topic or theme detection using this paradigm is yet to be realized.

IV. Research Issues in Automatic Speech Recognition and Understanding

The progress toward automatic speech recognition and understanding achieved in the past two decades is quite remarkable. There are currently many commercialization efforts that try to capitalize on the engineering advances as outlined above. There is, however, just as much desire to move forward with further research so that a machine that can truly converse with a human becomes possible. We have discussed in previous sections ongoing research issues associated with each individual technological component. We further discuss here several remaining research issues, mostly immediate problems that are ahead of us, in the hope that a definable technical direction would become obvious.

A. Robustness

The development of statistical methods, which make the system both easy to design, in terms of implementation, and capable of delivering somewhat sufficient performance, in limited tasks, has attracted enthusiasm in technology investment. However, one needs to be rather careful in understanding the permissible operating conditions under which deployment of the system is viable. These conditions include the level of background noise, channel distortion and its variation, speaker dependency, allowable speaking styles and syntactic deviation, spontaneity of the speech, and so on. At present, a system would fail to deliver satisfactory performance if it is not used within the intended, often very narrowly defined, operating condition. Compared to a human listener, most of the spoken language systems do not perform well when actual operating conditions deviate from the intended ones. This gives rise to the concern of the robustness of a spoken language system.

With the statistical method, which is data driven, one can in general improve the system performance by providing training data collected under the exact intended deployment condition. Although a system trained and operated under noisy conditions will still not perform as well as a system trained and operated in a quiet acoustic ambient, its performance will be substantially better than that of a system trained in a quiet but operated in a noisy (mismatched) condition. The problem is that collecting the "right data" (i.e., under a matched condition) is often very costly. This is particularly true in the case of voice-enabled services in the telephone network due to the diversity in network systems and telephone devices. (For example, Fig. 6 shows

the range of variation in the frequency response of telephone handsets.) The same notion applies to robustness against language models, speaking styles, as well as other conditions. For a voice-activated dictation machine serving a specific user for a limited domain of application, training the system to perform a particular task may not pose a serious logistic problem. For voice-activated services over the communication network for thousands of users, it is nontrivial to collect data to ensure coverage of the operating conditions. In other words, one should expect various degrees of condition "mismatch" between design/training in the laboratory and deployment in the field in almost all systems. The issue of robustness, thus, is to address the system's inherent capability in dealing with the mismatch conditions.

In the context of statistical pattern recognition, the mismatch means that the maximum *a posteriori* decision rule of (6) is being implemented as, with $\boldsymbol{Y}$ denoting the actually received signal

$$C(\boldsymbol{Y}) = C_i \qquad \text{if } P_\Lambda(C_i|\boldsymbol{Y}) = \max_j P_\Lambda(C_j|\boldsymbol{Y}) \tag{35}$$

although the parameter Λ has been obtained based on $\boldsymbol{X}$. In general, we assume $\boldsymbol{Y} = h(\boldsymbol{X}, \theta)$ defined on some unknown parameter θ. The approach to the robustness issue can, thus, be addressed in several ways. One is to find and use an invariant feature to represent the speech. An ideal invariant feature is a representation that will not fluctuate with the signal conditions; the same parameter Λ trained on $\boldsymbol{X}$ is expected to remain applicable for $\boldsymbol{Y}$. This is obviously difficult to achieve. Another rather prevalent approach is to embed the function h in the *a posteriori* probability, i.e., to use $P_\Lambda(C_i|\boldsymbol{X}) = P_\Lambda(C_i|h^{-1}(\boldsymbol{Y}, \theta))$ in the decision rule. The interference parameter θ sometimes can be estimated from $\boldsymbol{Y}$.

The most immediate concern in "condition mismatch" is noise and distortion. For additive noise, it is customary to assume, with $\boldsymbol{X}$ and $\boldsymbol{Y}$ being the "clean" and the "noisy" (observed) power spectral sequences, respectively

$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{N} \tag{36}$$

where $\boldsymbol{N}$ is the sequence of noise spectra of an unknown (and possibly varying) level. If the interference is a linear distortion, then

$$\boldsymbol{Y} = h(\boldsymbol{X}) = \boldsymbol{H}\boldsymbol{X} \tag{37}$$

where $\boldsymbol{H}$ is the frequency response of the linear distortion model. Note that in the case of linear distortion, (37) reduces to the form of (36) when a cepstral representation is used. The two types of interference are sometimes lumped together into a simplified function

$$\boldsymbol{Y} = h(\boldsymbol{X}) = \boldsymbol{H}\boldsymbol{X} + \boldsymbol{N}. \tag{38}$$

Much of the work toward robust speech recognition in the past decade focused on estimation of the parameters ($\boldsymbol{H}$ and $\boldsymbol{N}$) using $\boldsymbol{Y}$ [49]. Techniques such as spectral mean subtraction, signal bias removal, and maximum-likelihood linear regression fall in this category.

The robustness issue can also encompass normalization of the observation to compensate for the variation due to talker differences. One technique that attempts to normalize the spectral difference due to vocal tract length variation among talkers was shown to bring about small but consistent improvement in speech recognition accuracy.

Another thrust to enhance the robustness in system performance is the area of adaptation. Following the above formulation, adaptation is to find $P_{\Lambda'}(C_i|\boldsymbol{Y})$ from $P_\Lambda(C_i|\boldsymbol{X})$ based on a set of newly collected/observed data $\{\boldsymbol{Y}\}$ and some prior knowledge of the distribution of Λ. The technique is effective for converting the speech model (either speaker dependent or speaker independent) to that of another talker using a limited but reasonable amount of new data. Speaker normalization and adaptation techniques in a nonstatistical context have been an area of research for decades.

Adaptation techniques can also be useful for adapting the speech models to a new operating environment [50]. The paper by Lee and Huo [51] in this special issue attempts to address various aspects of the adaptation framework.

Designing a speech recognition and understanding system that works for a broad range of speaking styles and syntactic variability has not been as well understood. This is one critical area for research.

B. Language Structure and Representation

It is argued that an HMM with a mixture observation density in each state can adequately represent the acoustic variation manifested in the distribution of spectral parameters. This is due to the density approximation capability of such a model. Beyond the variation at the local acoustic level, however, the probabilistic nature of a language is often less understood. An expression of language is conventionally treated as an event governed by a set of prescribed rules rather than a random phenomenon. Setting aside the colloquial and pragmatic aspects of the language, we judge a (written) sentence to be either grammatical or ungrammatical, but never, say, 67% grammatically correct in an analytical sense.

The lack of a systematic study in probabilistically interpreting a language, as well as a large collection of statistical data, results in two technical areas in need of further research. One pertains to the representation of the linguistic structure ready for the application of probabilistic methods and the other the estimation method for reliable derivation of the relevant statistics for use in speech recognition and understanding. The former issue is equivalent to the definition of an event space based upon which a probabilistic model can be developed. Without such a representation, it is difficult to analyze the outcome of the statistical model.

Grammar is the rule that governs a language. In terms of language processing, the complexity is compounded by the interaction between the structural rules and the lexical elements of expression such as words and phrases. Traditionally, linguists establish a grammar (the structural rule) based on elementary classes such as noun, verb, adjective, noun phrase,

and so on, devoid of direct association of specific lexical element. However, the variation in our expression of message or concept comes from possibly three essential components: the choice of lexical elements (words and phrases), the grammatical structure (one may argue that it's less probabilistic), and the interaction between them. While developing a linguistic theory that encompasses all these elements is possible, it is not straightforward to address these elements of uncertainty and cast them in a formal probabilistic framework.

A number of grammatical representations and parsers exist [43]. The most pervasive is the finite state grammar (see Section III-E1), which provides an integrated mechanism for addressing both variations in the sentential structure (traditionally addressed by a parser) and the choice of words. It is, however, a simplified and crude model of language. An N-gram language model is a special case (fixed-order) finite state grammar: it addresses the probability of observing a word following a particular sequence of $N-1$ words. A finite state grammar such as an N-gram model has the advantage of implementational ease. The fundamental issue with a finite state grammar is the difficulty in having a precise coverage. Overspecification (which often happens with a high-order finite state machine) leads to frequent encounters of out-of-grammar but legitimate expressions (i.e., undercoverage of the overall linguistic expressions). Underspecification, on the contrary, will have overcoverage, which, while it alleviates out-of-grammar problems, reduces the accuracy and the effectiveness of the estimated probabilistic language model (e.g., many unlikely or impossible expressions in reality would have nonnegligible probability assignments).

The issue of representation for a sequence of linguistic events also exists at the lexical level. People pronounce words differently due to many reasons. The realized phonemic content of a phrase in a spoken utterance can vary rather vastly. The lexicon a system uses for decoding a word is usually insufficient. For example, as reported in a study sponsored by DARPA, 37 different pronunciations of the word "*the*" were found in a data set of fewer than 5000 word tokens. In the same data set, more than half of the word tokens were not "properly" pronounced according to a well-compiled dictionary based on several millions of spoken words. The range of pronunciation variation is enormous. The need for a pronunciation dictionary with a proper coverage to accommodate the variation is critical for a high-performance speech recognition system. The same question of structural representation applies here, although the implication of rules (grammatical versus lexical) is different. Research in this area in the past few years only produced slight improvements in recognition accuracy. The fundamental issue of a proper representation is still open.

Another open issue associated with language modeling is the criterion employed in parameter estimation. Traditionally, one uses the perplexity (31) or the entropy (30) estimated by the language model as a measure of quality for the model. Optimization in language modeling aims at minimizing the entropy associated with the resultant language model. While entropy is a measure that provides a framework for analysis of the information content of a system [42], its correlation with the performance of a speech recognition system is hardly proportional. In recent studies of large-vocabulary continuous spontaneous speech recognition (the Switch Board task of DARPA) employing either an N-gram or a tree-structured language model, it was found that reduction in perplexity or entropy does not translate into reduction in the recognition error rate. There are a number of reasons. For one, the entropy, used as an estimation criterion in language modeling, does not have a direct coupling with the entropy at the acoustic level. It is possible that a reduction in the language perplexity (note: as measured by the language model) may cause increase in acoustic perplexity (note: as measured by the acoustic model). The issue of structural representation as well as the coverage problem further makes the optimization result difficult to interpret in terms of its true performance. The problem in structural representation and statistical estimation is even more serious when it comes to spoken utterances as opposed to written texts. In spoken utterances, ill-formed sentences with disfluencies such as repair, partial, and repetitive words are more often observed than otherwise. These ill-formed, as well as many other colloquial sentences, obviously deviate from the grammatical rules and usually lack the regularity to bring about a statistical significance. Language modeling for spoken utterances, from the structural representation to model adaptation to a particular talker (people's speaking habits differ), is one of the major challenges in this field of research.

Adaptation or acquisition of the language structure to a particular communication context is also a worthwhile and active area of research.

C. Database and Generalization

As discussed above, data-driven methods have brought about fruitful results in the past decade. Unlike the traditional approach, in which knowledge of the speech behavior is "discovered" and "documented" by human experts, statistical methods provide an automatic procedure to "learn" the regularities in the speech data directly. The need of a large set of good training data is, thus, more critical than ever.

Establishing a good speech database for the machine to uncover the characteristics of the signal is not trivial. There are basically two broad issues to be carefully considered: one being the content and its annotation, and the other the collecting mechanism.

The content of a database must reflect the intended use of the database. For simple command and control applications, this is relatively straightforward: the data collected must contain all the command words. For general dictation applications, the data collected for training the acoustic unit models may not be quite the same as the data for training the language model, mainly because the effort may be unmanageable. In order to train a reasonably good general-English language model, a text database on the order of 100 million words is needed. When spoken, this would amount to over 10 000 hours of speech, which no single talker can accomplish. By separating the text and the acoustic aspects of the

database, one alleviates the effort problem but needs to address the issue of estimation consistency [52]. For speaker-independent applications, particularly those deployed over the telephone network, the data-collection process can be very involved, as the system designer needs to consider the range of regional dialects and accents.

For natural dialogue applications such as the Air Travel Information System in the DARPA program [45], a wizard setup is often used to collect the data. A wizard in this case is a human mimicking the machine in interacting with the user. Through the interaction, natural queries in sentential forms are collected. A committee is called upon to resolve cases that may be ambiguous in certain aspects. While a wizard setup can produce a useful set of data, it lacks the diversity, particularly in situations where the real machine may fail. A human wizard cannot intentionally simulate a machine in error, and, thus, the recorded data fail to provide information of real human–machine interaction.

The recorded data need to be verified, labeled, and annotated by people whose knowledge will be introduced into the design of the system through this learning process (i.e., via supervised training of the system after the data have been labeled). Labeling and annotation for isolated word utterances may be straightforward but tedious when the amount of data is large. For continuous speech recognition and understanding, nevertheless, this process can easily become unmanageable. For example, how do we annotate speech repairs and partial words, how do the phonetic transcribers reach a consensus in acoustic-phonetic labels when there is ambiguity, and how do we represent a semantic notion? Errors in labeling and annotation will result in system performance degradation. How to ensure the quality of the annotated results is, thus, of major concern. Research in automating or creating tools to assist the verification procedure is by itself an interesting subject.

The data-collection mechanism has to be attended with care. For example, a system deployed for digital cellular phone users needs to take into account the speech coder characteristics. But, an improperly designed antialiasing filter for analog-to-digital conversion at the front end of the system should not be considered part of the adverse effect or source of variability and needs to be corrected before data collection can begin. The confusion between recorded diversity in operating conditions and the unwanted interference or adverse effects due to misuse of equipment often exists. An unwanted interference will cause detriment in the training result, but a true coverage of the diverse operating condition (e.g., real ambient noise) is crucial in guaranteeing a satisfactory performance. Thorough understanding and examination of the signal is very important.

Another area of research that has gained interest is a modeling methodology and the associated data-collection scheme that can reduce the task dependency. To maximize the performance, one should always strive for data that truly reflects the operating condition. It, thus, calls for a database collection plan that is consistent with the task. This data-collection effort would soon become unmanageable if the system designer has to redo data collection for each and every application that is being developed. It is, therefore, desirable to design a task-independent data set and a modeling method that delivers a reasonable performance upon first use and can quickly allow in-field trials for further revision as soon as task-dependent data become available. Research results in this area can offer the benefit of a reduced application development cost.

D. Human–Machine Dialogue

Human–machine interaction certainly has various levels of complexity. We have laid out several elementary steps. The simplest is an isolated word recognizer that "understands" a human's command. A speech recognizer operating on word-spotting attempts to detect a human's command or intention by focusing on the key words or phrases embedded in the stream of sounds, likely a natural sentence, from the talker. A more sophisticated system can also ask the user to fill in a set of prescribed information fields (e.g., "your birthday please?" or "your credit-card number please?") in order to derive the needed action. Solicitation of information can be cast in a natural inquisitive form such as the example above, or as an option menu ("The following services are available. . . . Which do you like?"). This type of interaction is, strictly speaking, a one-way communication and in the context of human–machine dialogue is often referred to as a system-driven dialogue.

A true dialogue goes beyond an interrogation for information (either from the user to the machine or vice versa). A dialogue involves response based on the previous state of the conversation. Both the user and the machine should be able to ask for clarification or extension to the previous information exchange, or to initiate a new domain of interaction. This more natural form of dialogue is referred to as a mixed initiative or variable initiative dialogue.

Research issues in this area can be divided into two broad categories. One involves artificial intelligence because, in order for a machine to be able to perform natural dialogue, it must have acquired an ability to communicate (as opposed to just recognize). Such ability is far beyond speech recognition and understanding; it must have a store of "knowledge" and "content," not just the protocol or handling of the communication mechanism, to support the communication. Knowledge representation and retrieval, database organization and search, semantic inference, and decision support are all required to various degrees. This is obviously a very broad and long-term challenge.

The other research need is much narrower but immediate. Before it is possible to design a machine that can communicate, it is often desirable to provide system design tools to allow human intervention, either at the application design stage or during run-time. Tools that allow the human–machine interaction system designer to develop a task based on his or her anticipation or envisage of the system behavior in response to a majority of users are very useful. For example, tools that provide efficient design of dialogue states and flow, and allow the developer to revise and support the operation of the system are deemed extremely valuable.

V. Conclusion

Over three decades of research in spoken language processing have produced remarkable advances in automatic speech recognition and understanding that helps us take a big step toward natural human–machine communication. Signal-processing techniques led to a better understanding of speech characteristics, providing deep insights into acoustic-phonetic properties of a language. The introduction of a statistical framework not only makes the problem of automatic recognition of speech tractable but also paves the road to practical engineering system designs. It was found that a particular probabilistic measure, the HMM, provides a speech modeling formalism that is powerful and yet easy to implement. Coupled with a finite state representation of a language, hidden Markov modeling has become the underpinning of most of today's speech-recognition and understanding systems under deployment. To accomplish the ultimate goal of a machine that can communicate with people, however, a number of research issues are awaiting further study. Such a communicating machine needs to be able to deliver a satisfactory performance under a broad range of operating conditions and have an efficient way of representing, storing, and retrieving "knowledge" required in a natural conversation. With the current enthusiasm in research advances, we are optimistic that the Holy Grail of natural human–machine communication will soon be within our technological reach.

References

[1] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," *J. Acoust. Soc. Amer.*, vol. 24, no. 6, pp. 637–642, 1952.

[2] H. F. Olson and H. Belar, "Phonetic typewriter," *J. Acoust. Soc. Amer.*, vol. 28, no. 6, pp. 1072–1081, 1956.

[3] J. W. Forgie and C. D. Forgie, "Results obtained from a vowel recognition computer program," *J. Acoust. Soc. Amer.*, vol. 31, no. 11, pp. 1480–1489, 1959.

[4] J. Suzuki and K. Nakata, "Recognition of Japanese vowels—Preliminary to the recognition of speech," *J. Radio Res. Lab.*, vol. 37, no. 8, pp. 193–212, 1961.

[5] T. Sakai and S. Doshita, "The phonetic typewriter, information processing 1962," presented at the *Proc. IFIP Congr.*, Munich, Germany, 1962.

[6] K. Nagata, Y. Kato, and S. Chiba, "Spoken digit recognizer for Japanese language," *NEC Res. Develop.*, no. 6, 1963.

[7] D. B. Fry, "Theoretical aspects of the mechanical speech recognition," *J. Br. Inst. Radio Eng.*, vol. 19, no. 4, pp. 211–229, 1959.

[8] T. B. Martin, A. L. Nelson, and H. J. Zadell, "Speech recognition by feature abstraction techniques," Air Force Avionics Lab, Tech. Rep. AL-TDR-64-176, 1964.

[9] T. K. Vintsyuk, "Speech discrimination by dynamic programming," *Kibernetika*, vol. 4, pp. 81–88, Jan.–Feb. 1968.

[10] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 43–49, Feb. 1978.

[11] F. Jelinek, L. R. Bahl, and R. L. Mercer, "Design of a linguistic statistical decoder for the recognition of continuous speech," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 250–256, 1975.

[12] F. Jelinek, "The development of an experimental discrete dictation recognizer," in *Proc. IEEE*, vol. 73, Nov. 1985, pp. 1616–1624.

[13] L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon, "Speaker independent recognition of isolated words using clustering techniques," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 336–349, Aug. 1979.

[14] F. Itakura, "Minimum prediction residual applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 67–72, Feb. 1975.

[15] B. H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass liftering in speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 947–954, July 1987.

[16] J. G. Wilpon and L. R. Rabiner, "A modified K-means clustering algorithm for use in isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 587–594, June 1985.

[17] J. G. Wilpon, L. R. Rabiner, C. H. Lee, and E. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 1870–1878, Nov. 1990.

[18] D. R. Reddy, "An approach to computer speech recognition by direct analysis of the speech wave," Comput. Sci. Dept., Stanford Univ., Tech. Rep. C549, Sept. 1966.

[19] V. R. Lesser, R. D. Fennell, L. D. Erman, and D. R. Reddy, "Organization of the hearsay—II: Speech understanding system," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 11–23, June 1975.

[20] J. Ferguson, Ed., *Hidden Markov Models for Speech*. Princeton, NJ: IDA, 1980.

[21] B. H. Juang, "Maximum likelihood estimation for mixture multivariate stochastic observations of Markov chains," *AT&T Tech. J.*, vol. 64, 1985.

[22] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proc. IEEE*, vol. 77, Feb. 1989, pp. 257–286.

[23] R. Pieraccini and E. Levin, "Stochastic representation of semantic structure for speech understanding," in *Proc. Eurospeech 91*, Genova, Italy, Sept. 1991.

[24] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*. New York: Springer-Verlag, 1972.

[25] H. C. Wang, M.-S. Chen, and T. Yang, "A novel approach to the speaker identification over telephone networks," in *Proc. ICASSP-93*, Minneapolis, MN, Apr. 1993, vol. 2, pp. 407–410.

[26] R. V. Cox, *et al.*, "Speech and language processing for next-millenium communications services," *Proc. IEEE*, vol. 88, pp. 1314–1337, Aug. 2000.

[27] B. H. Juang, "Automatic speech recognition: Problems, progress & prospects," presented at the IEEE Workshop Neural Networks for Signal Processing, Kyoto, Japan, Oct. 1996.

[28] D. Pallett, *et al.*, "DARPA HUB-4 rep.," National Institute of Science and Technology, Feb. 1999.

[29] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[30] A. Andreou, T. Kamm, and J. Cohen, "Experiments in vocal tract normalization," presented at the *Proc. CAIP/Rutgers Workshop: Frontiers in Speech Recognition II*, 1994.

[31] J. D. Markel and A. H. Gray Jr., *Linear Prediction of Speech*. Berlin, Germany: Springer-Verlag, 1976.

[32] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[33] C. Nadeu, P. Paches-Leal, and B. H. Juang, "Filtering the time sequences of spectral parameters for speech recognition," *Speech Commun.*, vol. 22, pp. 315–332, 1997.

[34] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 578–589, Oct. 1994.

[35] 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, Feb. 1986.

[36] ——, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 254–272, Apr. 1981.

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

[38] L. E. Baum, T. Petri, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Statist.*, vol. 41, pp. 164–171, 1970.

[39] B. H. Juang and S. Katagiri, "Discriminative training," *J. Acoust. Soc. Jpn (E)*, vol. 13, no. 6, pp. 333–339, 1992.

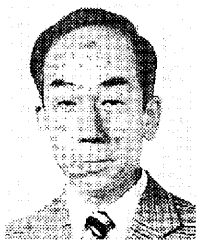
[40] B. H. Juang, W. Chou, and C. H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 257–265, May 1997.

[41] B. H. Juang and L. R. Rabiner, "A probabilistic distance measure for hidden Markov models," *AT&T Tech. J.*, vol. 64, pp. 391–408, Feb. 1985.
[42] F. Jelinek, *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press, 1997.
[43] R. C. Moore, "Using natural-language knowledge sources in speech recognition," in *Computational Models of Speech Pattern Processing*, K. Ponting, Ed. Berlin, Germany: Springer-Verlag, 1997, pp. 304–327.
[44] H. Ney and S. Ortmanns, "Progress in dynamic programming search for LVCSR," *Proc. IEEE*, vol. 88, pp. 1224–1240, Aug. 2000.
[45] "ATIS Tech. Rep.," in *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, 1995, pp. 241–280.
[46] L. G. Miller and A. Gorin, "Structured networks for adaptive language acquisition," *Int. J. Pattern Recognit. Artif. Intell. (Special Issue on Neural Networks)*, vol. 7, no. 4, pp. 873–898, 1993.
[47] S. Deerwester, *et al.*, "Indexing by latent semantic analysis," *J. Amer. Soc. Inform. Sci.*, vol. 41, pp. 391–407, 1990.
[48] T. Kawahara, C. H. Lee, and B. H. Juang, "Combining key-phrase detection and subword based verification for flexible speech understanding," in *Proc. IEEE ICASSP97*, May 1997.
[49] J.-C. Junqua and J.-P. Haton, *Robustness in Automatic Speech Recognition*. Boston, MA: Kluwer, 1996.
[50] S. Furui, "Recent advances in robust speech recognition," in *Proc. ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mouson, France, 1997, pp. 11–20.
[51] C. H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," *Proc. IEEE*, vol. 88, pp. 1241–1269, Aug. 2000.
[52] A. Nadas, D. Nahamoo, and M. A. Picheny, "On a model-robust training method for speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1432–1436, 1988.
[53] P. Denes, "The design and operation of the mechanical speech recognizer at University College London," *J. Br. Inst. Radio Eng.*, vol. 19, no. 4, pp. 211–229, 1959.



Biing-Hwang (Fred) Juang (Fellow, IEEE) is Director of Acoustics and Speech Research at Bell Labs, Lucent Technologies, Murray Hill, NJ. He is engaged in a wide range of communication-related research activities, from speech coding and speech recognition to multimedia communications. He has published extensively and holds a number of patents in the areas of speech communication and communication services. He is coauthor of the book *Fundamentals of Speech Recognition* (Englewood Cliffs, NJ: Prentice-Hall).

Dr. Juang received the 1993 Best Senior Paper Award, the 1994 Best Senior Paper Award, and the 1994 Best Signal Processing Magazine Paper Award, and was coauthor of a paper granted the 1994 Best Junior Paper Award, all from the IEEE Signal Processing Society. In 1997, he won the Bell Labs' President Award for leading the Bell Labs Automatic Speech Recognition (BLASR) team. He also received the prestigious 1998 Signal Processing Society's Technical Achievement Award and was named the Society's 1999 Distinguished Lecturer. In 2000, he was awarded the IEEE Third Millennium Medal for his contributions to the field of speech processing and communications. He was also named 1999 Bell Labs Fellow. He was an Editor for the IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING (1986–88), the IEEE TRANSACTIONS ON NEURAL NETWORKS (1992–93), and the *Journal of Speech Communication* (1992–94). He has served on the Digital Signal Processing and Speech Technical committees as well as the Conference Board of the IEEE Signal Processing Society and was 1991–1993 Chairman of the Technical Committee on Neural Networks for Signal Processing. He is currently Editor-in-Chief of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING and Member of Editorial Board of the PROCEEDINGS OF THE IEEE. He also serves on international advisory boards outside the United States.



Sadaoki Furui (Fellow, IEEE) is currently both a Professor at the Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan, and an Instructor at Tokyo University, Tokyo. From 1978 to 1979, he was a Member of Staff of the Acoustics Research Department, Bell Laboratories, Murray Hill, NJ, as a Visiting Researcher working on speaker verification. He is engaged in a wide range of research on speech analysis, speech recognition, speaker recognition, and speech synthesis, and has authored or coauthored more than 300 published articles. He is the author of *Digital Speech Processing, Synthesis, and Recognition* (New York: Marcel Dekker, 1989) in English, *Digital Speech Processing* (Tokyo, Japan: Tokai University Press, 1985) in Japanese, *Acoustics and Speech Processing* (Japan: Kindai-Kagaku-Sha, 1992) in Japanese, and *Speech Information Processing* (Japan: Morikita, 1998) in Japanese. He edited *Advances in Speech Signal Processing* (New York: Marcel Dekker, 1992) jointly with Dr. M. M. Sondhi. He translated into Japanese *Fundamentals of Speech Recognition*, authored by Dr. L. R. Rabiner and Dr. B.-H. Juang (NTT Advanced Technology, 1995), and *Vector Quantization and Signal Compression*, authored by Dr. A. Gersho and Dr. R. M. Gray (Corona-sha, 1998).

Dr. Furui received the Yonezawa Prize and the Paper Award from the Institute of Electronics, Information and Communication Engineers of Japan (IEICE) (1975, 1988, 1993), and the Sato Paper Awards from the Acoustical Society of Japan (ASJ) (1985, 1987). He received the Senior Award from the IEEE ASSP Society (1989) and the Achievement Award from the Minister of Science and Technology, Japan (1989). He also received the Book Award from the IEICE (1990). He is a Fellow of the Acoustical Society of America. From 1995 to 1997, he served as a Vice President of the Acoustical Society of Japan (ASJ). He served on the IEEE Technical Committee on Speech and MMSP and on numerous IEEE conference organizing committees. He is a Board member of both the International Speech Communication Association and the ASJ, and Vice President of the Permanent Council for International Conferences on Spoken Language Processing. He is an Editor-in-Chief of the *Journal of Speech Communication* and Member of Editorial Board of the *Journal of Computer Speech and Language*. In 1993, he served as an IEEE SPS Distinguished Lecturer.