Survey of the State of the Art in Human Language Technology

Sponsors:
National Science Foundation
Directorate XIII-E of the Commission of the European Communities
Center for Spoken Language Understanding, Oregon Graduate Institute

November 21, 1995

Contents

Chapter 1

Spoken Language Input

1.1 Overview

Victor Zue[a] & Ron Cole[b]

[a] MIT Laboratory for Computer Science, Cambridge, Massachusetts, USA
[b] Oregon Graduate Institute of Science & Technology, Portland, Oregon, USA

Spoken language interfaces to computers is a topic that has lured and fascinated engineers and speech scientists alike for over five decades. For many, the ability to converse freely with a machine represents the ultimate challenge to our understanding of the production and perception processes involved in human speech communication. In addition to being a provocative topic, spoken language interfaces are fast becoming a necessity. In the near future, interactive networks will provide easy access to a wealth of information and services that will fundamentally affect how people work, play and conduct their daily affairs. Today, such networks are limited to people who can read and have access to computers—a relatively small part of the population even in the most developed countries. Advances in human language technology are needed for the average citizen to communicate with networks using natural communication skills using everyday devices, such as telephones and televisions. Without fundamental advances in user-centered interfaces, a large portion of society will be prevented from participating in the age of information, resulting in further stratification of society and tragic loss in human potential.

The first chapter in this survey deals with spoken language *input* technologies. A speech interface, in a user's own language, is ideal because it is the most natural, flexible, efficient, and economical form of human communication. The following sections summarize spoken input technologies that will facilitate such an interface.

Spoken input to computers embodies many different technologies and applications, as shown in Figure 1.1. In some cases, as shown at the bottom of the figure, one is interested not in the underlying linguistic content, but the identity of the speaker, or the language being spoken. Speaker recognition can involve *identifying* a specific speaker out of a known population, which has forensic implications, or *verifying* the claimed identity of a user, thus enabling controlled access to locales (e.g., a computer room) and services (e.g., voice banking). Speaker recognition technologies are addressed in section 1.7. Language identification also has important applications, and techniques applied to this area are summarized in section 8.7.

When one thinks about speaking to computers, the first image is usually speech recognition, the conversion of an acoustic signal to a stream of words. After many years of research, speech recognition technology is beginning to pass the threshold of practicality. The last decade has witnessed dramatic improvement in speech recognition technology, to the extent that high performance algorithms and systems are becoming available. In some cases, the transition from laboratory demonstration to commericial deployment has already begun. Speech input capabilities are emerging that can provide functions like voice dialing (e.g., *Call home*), call routing (e.g., *I would like to make a collect call*), simple data entry (e.g., entering a credit card number), and preparation of structured documents (e.g., a radiology report). The basic issues of speech recognition, together with a summary of the state-of-the-art, is described in section 1.2. As these authors point out, speech recognition involves several component technologies. First, the digitized signal must be transformed into a set of measurements. This *signal representation* issue is elaborated in section 1.3. Section 1.4 discusses techniques that enable the system to achieve robustness in the presence of transducer and environmental variations, and techniques for adapting to these variations. Next, the various speech sounds must be modeled appropriately. The most widespread technique for acoustic modeling is called hidden Markov modeling (HMM), and is the subject of section 1.5. The search for the final answer involves the use of language constraints, which is covered in section 1.6.

Speech recognition is a very challenging problem in its own right, with a well defined set of applications. However, many tasks that lend themselves to spoken input—making travel arrangements or selecting a movie—are in fact exercises in interactive problem solving. The solution is often built up incrementally, with both the user and the computer playing active roles in the "conversation." Therefore, several language-based input and output technologies must be developed and integrated to reach this goal. The remainer of Figure 1.1 shows the major components of a typical conversational system. The spoken input is first processed through the speech recognition component. The natural language component, working in concert with the recognizer, produces a meaning representation. The final section of this chapter, on spoken language

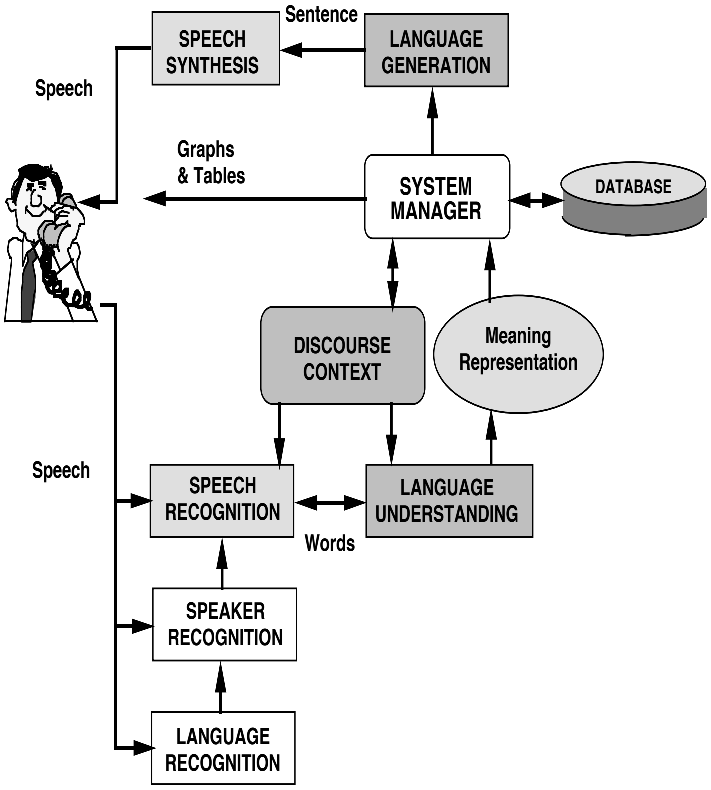

Figure 1.1: Technologies for spoken language interfaces.

understanding technology (section 1.8), discusses the integration of speech recognition and natural language processing techniques.

For information retrieval applications illustrated in this figure, the meaning representation can be used to retrieve the appropriate information in the form of text, tables and graphics. If the information in the utterance is insufficient or ambiguous, the system may choose to query the user for clarification. Natural language generation and speech synthesis, covered in chapters 4 and 5, respectively, can be used to produce spoken responses that may serve to clarify the tabular information. Throughout the process, discourse information is maintained and fed back to the speech recognition and language understanding components, so that sentences can be properly understood in context.

1.2 Speech Recognition

Victor Zue,[a] Ron Cole,[b] & Wayne Ward[c]

[a] MIT Laboratory for Computer Science, Cambridge, Massachusetts, USA
[b] Oregon Graduate Institute of Science & Technology, Portland, Oregon, USA
[c] Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

1.2.1 Defining the Problem

Speech recognition is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. The recognized words can be the final results, as for applications such as commands & control, data entry, and document preparation. They can also serve as the input to further linguistic processing in order to achieve speech understanding, a subject covered in section 1.8.

Speech recognition systems can be characterized by many parameters, some of the more important of which are shown in Figure 1.1. An isolated-word speech recognition system requires that the speaker pause briefly between words, whereas a continuous speech recognition system does not. Spontaneous, or extemporaneously generated, speech contains disfluencies, and is much more difficult to recognize than speech read from script. Some systems require speaker enrollment—a user must provide samples of his or her speech before using them, whereas other systems are said to be speaker-independent, in that no enrollment is necessary. Some of the other parameters depend on the specific task. Recognition is generally more difficult when vocabularies are large or have many similar-sounding words. When speech is produced in a sequence of words, language models or artificial grammars are used to restrict the combination of words. The simplest language model can be specified as a finite-state network, where the permissible words following each word are given explicitly. More general language models approximating natural language are specified in terms of a context-sensitive grammar.

One popular measure of the difficulty of the task, combining the vocabulary size and the language model, is *perplexity*, loosely defined as the geometric mean of the number of words that can follow a word after the language model has been applied (see section 1.6 for a discussion of language modeling in general and perplexity in particular). Finally, there are some external parameters that can affect speech recognition system performance, including the characteristics of the environmental noise and the type and the placement of the microphone.

Speech recognition is a difficult problem, largely because of the many sources of variability associated with the signal. First, the acoustic realizations of phonemes, the

Parameters	**Range**
Speaking Mode	Isolated words to continuous speech
Speaking Style	Read speech to spontaneous speech
Enrollment	Speaker-dependent to Speaker-independent
Vocabulary	Small ($<$ 20 words) to large ($>$ 20,000 words)
Language Model	Finite-state to context-sensitive
Perplexity	Small ($<$ 10) to large ($>$ 100)
SNR	High ($>$ 30 dB) to low ($<$ 10 dB)
Transducer	Voice-cancelling microphone to telephone

Table 1.1: Typical parameters used to characterize the capability of speech recognition systems

smallest sound units of which words are composed, are highly dependent on the context in which they appear. These *phonetic variabilities* are exemplified by the acoustic differences of the phoneme[1] /t/ in *two, true,* and *butter* in American English. At word boundaries, contextual variations can be quite dramatic—making *gas shortage* sound like *gash shortage* in American English, and *devo andare* sound like *devandare* in Italian.

Second, *acoustic variabilities* can result from changes in the environment as well as in the position and characteristics of the transducer. Third, *within-speaker variabilities* can result from changes in the speaker's physical and emotional state, speaking rate, or voice quality. Finally, differences in sociolinguistic background, dialect, and vocal tract size and shape can contribute to *across-speaker variabilities.*

Figure 1.2 shows the major components of a typical speech recognition system. The digitized speech signal is first transformed into a set of useful measurements or features at a fixed rate, typically once every 10–20 msec (see sections 1.3 and 11.3 for signal representation and digital signal processing, respectively). These measurements are then used to search for the most likely word candidate, making use of constraints imposed by the acoustic, lexical, and language models. Throughout this process, training data are used to determine the values of the model parameters.

Speech recognition systems attempt to model the sources of variability described above in several ways. At the level of signal representation, researchers have developed representations that emphasize perceptually important speaker-independent features of

[1]Linguistic symbols presented between slashes, e.g., /p/, /t/, /k/, refer to *phonemes*; the minimal sound unit by changing it one changes the meaning of a word. The acoustic realizations of phonemes in speech are referred to as *allophones*, *phones*, or *phonetic segments*, and are presented in brackets, e.g., [p], [t], [k].

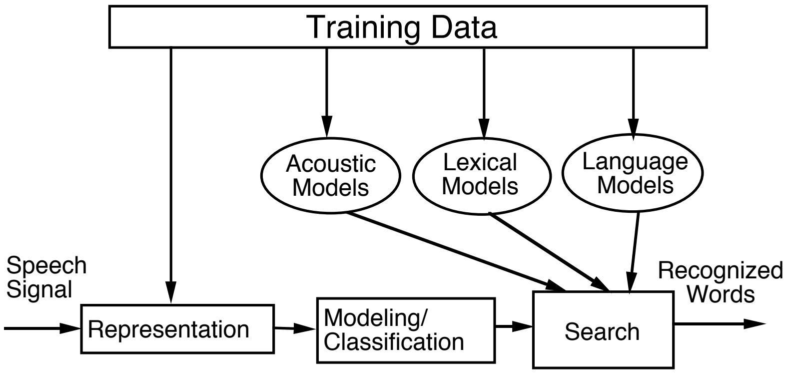

Figure 1.2: Components of a typical speech recognition system.

the signal, and de-emphasize speaker-dependent characteristics (Hermansky, 1990). At the acoustic phonetic level, speaker variability is typically modeled using statistical techniques applied to large amounts of data. Speaker adaptation algorithms have also been developed that adapt speaker-independent acoustic models to those of the current speaker during system use, (see section 1.4). Effects of linguistic context at the acoustic phonetic level are typically handled by training separate models for phonemes in different contexts; this is called context dependent acoustic modeling.

Word level variability can be handled by allowing alternate pronunciations of words in representations known as *pronunciation networks*. Common alternate pronunciations of words, as well as effects of dialect and accent are handled by allowing search algorithms to find alternate paths of phonemes through these networks. Statistical language models, based on estimates of the frequency of occurrence of word sequences, are often used to guide the search through the most probable sequence of words.

The dominant recognition paradigm in the past fifteen years is known as hidden Markov models (HMM). An HMM is a doubly stochastic model, in which the generation of the underlying phoneme string and the frame-by-frame, surface acoustic realizations are *both* represented probabilistically as Markov processes, as discussed in sections 1.5, 1.6 and 11.2. Neural networks have also been used to estimate the frame based scores; these scores are then integrated into HMM-based system architectures, in what has come to be known as *hybrid systems*, as described in section 11.5.

An interesting feature of frame-based HMM systems is that speech segments are identified during the search process, rather than explicitly. An alternate approach is to first identify speech segments, then classify the segments and use the segment scores to recognize words. This approach has produced competitive recognition performance in several tasks (Zue, Glass, et al., 1990; Fanty, Barnard, et al., 1995).

1.2.2 State of the Art

Comments about the state-of-the-art need to be made in the context of specific applications which reflect the constraints on the task. Moreover, different technologies are sometimes appropriate for different tasks. For example, when the vocabulary is small, the entire word can be modeled as a single unit. Such an approach is not practical for large vocabularies, where word models must be built up from subword units.

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

$$E = \frac{S + I + D}{N} 100$$

where N is the total number of words in the test set, and S, I, and D are the total number of substitutions, insertions, and deletions, respectively.

The past decade has witnessed significant progress in speech recognition technology. Word error rates continue to drop by a factor of 2 every two years. Substantial progress has been made in the basic technology, leading to the lowering of barriers to speaker independence, continuous speech, and large vocabularies. There are several factors that have contributed to this rapid progress. First, there is the coming of age of the HMM. HMM is powerful in that, with the availability of training data, the parameters of the model can be trained automatically to give optimal performance.

Second, much effort has gone into the development of large speech corpora for system development, training, and testing. Some of these corpora are designed for acoustic phonetic research, while others are highly task specific. Nowadays, it is not uncommon to have tens of thousands of sentences available for system training and testing. These corpora permit researchers to quantify the acoustic cues important for phonetic contrasts and to determine parameters of the recognizers in a statistically meaningful way. While many of these corpora (e.g., TIMIT, RM, ATIS, and WSJ; see section 12.3) were originally collected under the sponsorship of the U.S. Defense Advanced Research Projects Agency (ARPA) to spur human language technology development among its contractors, they have nevertheless gained world-wide acceptance (e.g., in Canada, France, Germany, Japan, and the U.K.) as standards on which to evaluate speech recognition.

Third, progress has been brought about by the establishment of standards for performance evaluation. Only a decade ago, researchers trained and tested their systems using locally collected data, and had not been very careful in delineating training and testing sets. As a result, it was very difficult to compare performance across systems, and a system's performance typically degraded when it was presented with previously unseen data. The recent availability of a large body of data in the public domain, coupled with the specification of evaluation standards, has resulted in uniform documentation of test results, thus contributing to greater reliability in monitoring progress (corpus development activities and evaluation methodologies are summarized in chapters 12 and 13 respectively).

Finally, advances in computer technology have also indirectly influenced our progress. The availability of fast computers with inexpensive mass storage capabilities has enabled researchers to run many large scale experiments in a short amount of time. This means that the elapsed time between an idea and its implementation and evaluation is greatly reduced. In fact, speech recognition systems with reasonable performance can now run in real time using high-end workstations without additional hardware—a feat unimaginable only a few years ago.

One of the most popular, and potentially most useful tasks with low perplexity ($PP = 11$) is the recognition of digits. For American English, speaker-independent recognition of digit strings spoken continuously and restricted to telephone bandwidth can achieve an error rate of 0.3% when the string length is known.

One of the best known moderate-perplexity tasks is the 1,000-word so-called Resource Management (RM) task, in which inquiries can be made concerning various naval vessels in the Pacific ocean. The best speaker-independent performance on the RM task is less than 4%, using a word-pair language model that constrains the possible words following a given word ($PP = 60$). More recently, researchers have begun to address the issue of recognizing spontaneously generated speech. For example, in the Air Travel Information Service (ATIS) domain, word error rates of less than 3% has been reported for a vocabulary of nearly 2,000 words and a bigram language model with a perplexity of around 15.

High perplexity tasks with a vocabulary of thousands of words are intended primarily for the dictation application. After working on isolated-word, speaker-dependent systems for many years, the community has since 1992 moved towards very-large-vocabulary (20,000 words and more), high-perplexity ($PP \approx 200$), speaker-independent, continuous speech recognition. The best system in 1994 achieved an error rate of 7.2% on read sentences drawn from North America business news (Pallett, Fiscus, et al., 1994).

With the steady improvements in speech recognition performance, systems are now being deployed within telephone and cellular networks in many countries. Within the next few years, speech recognition will be pervasive in telephone networks around the world. There are tremendous forces driving the development of the technology; in many countries, touch tone penetration is low, and voice is the only option for controlling automated services. In voice dialing, for example, users can dial 10–20 telephone numbers by voice (e.g., *call home*) after having enrolled their voices by saying the words associated with telephone numbers. AT&T, on the other hand, has installed a call routing system using speaker-*independent* word-spotting technology that can detect a few key phrases (e.g., *person to person*, *calling card*) in sentences such as: *I want to charge it to my calling card.*

At present, several very large vocabulary dictation systems are available for document generation. These systems generally require speakers to pause between words. Their performance can be further enhanced if one can apply constraints of the specific domain such as dictating medical reports.

Even though much progress is being made, machines are a long way from recognizing conversational speech. Word recognition rates on telephone conversations in the *Switchboard* corpus are around 50% (Cohen, Gish, et al., 1994). It will be many years before unlimited vocabulary, speaker-independent continuous dictation capability is realized.

1.2.3 Future Directions

In 1992, the U.S. National Science Foundation sponsored a workshop to identify the key research challenges in the area of human language technology, and the infrastructure needed to support the work. The key research challenges are summarized in Cole, Hirschman, et al. (1992). Research in the following areas for speech recognition were identified:

Robustness: In a robust system, performance degrades gracefully (rather than catastrophically) as conditions become more different from those under which it was trained. Differences in channel characteristics and acoustic environment should receive particular attention.

Portability: Portability refers to the goal of rapidly designing, developing and deploying systems for new applications. At present, systems tend to suffer significant degradation when moved to a new task. In order to return to peak performance, they must be trained on examples specific to the new task, which is time consuming and expensive.

Adaptation: How can systems continuously adapt to changing conditions (new speakers, microphone, task, etc) and improve through use? Such adaptation can occur at many levels in systems, subword models, word pronunciations, language models, etc.

Language Modeling: Current systems use statistical language models to help reduce the search space and resolve acoustic ambiguity. As vocabulary size grows and other constraints are relaxed to create more habitable systems, it will be increasingly important to get as much constraint as possible from language models; perhaps incorporating syntactic and semantic constraints that cannot be captured by purely statistical models.

Confidence Measures: Most speech recognition systems assign scores to hypotheses for the purpose of rank ordering them. These scores do not provide a good indication of whether a hypothesis is correct or not, just that it is better than the other hypotheses. As we move to tasks that require actions, we need better methods to evaluate the absolute correctness of hypotheses.

Out-of-Vocabulary Words: Systems are designed for use with a particular set of words, but system users may not know exactly which words are in the system vocabulary. This leads to a certain percentage of out-of-vocabulary words in natural conditions. Systems must have some method of detecting such out-of-vocabulary words, or they will end up mapping a word from the vocabulary onto the unknown word, causing an error.

Spontaneous Speech: Systems that are deployed for real use must deal with a variety of spontaneous speech phenomena, such as filled pauses, false starts, hesitations, ungrammatical constructions and other common behaviors not found in read speech. Development on the ATIS task has resulted in progress in this area, but much work remains to be done.

Prosody: Prosody refers to acoustic structure that extends over several segments or words. Stress, intonation, and rhythm convey important information for word recognition and the user's intentions (e.g., sarcasm, anger). Current systems do not capture prosodic structure. How to integrate prosodic information into the recognition architecture is a critical question that has not yet been answered.

Modeling Dynamics: Systems assume a sequence of input frames which are treated as if they were independent. But it is known that perceptual cues for words and phonemes require the integration of features that reflect the movements of the articulators, which are dynamic in nature. How to model dynamics and incorporate this information into recognition systems is an unsolved problem.

1.3 Signal Representation

Melvyn J. Hunt

Dragon Systems UK Ltd., Cheltenham, UK

In statistically based automatic speech recognition, the speech waveform is sampled at a rate between 6.6 kHz and 20 kHz and processed to produce a new representation as a sequence of vectors containing values of what are generally called *parameters*. The vectors ($y(t)$ in the notation used in section 1.5) typically comprise between 10 and 20 parameters, and are usually computed every 10 or 20 msec. These parameter values are then used in succeeding stages in the estimation of the probability that the portion of waveform just analyzed corresponds to a particular phonetic event that occurs in the phone-sized or whole-word reference unit being hypothesized. In practice, the representation and the probability estimation interact strongly: what one person sees as part of the representation another may see as part of the probability estimation process. For most systems, though, we can apply the criterion that if a process is applied to all speech it is part of the representation, while if its application is contingent on the phonetic hypothesis being tested it is part of the later matching stage.

Representations aim to preserve the information needed to determine the phonetic identity of a portion of speech while being as impervious as possible to factors such as speaker differences, effects introduced by communications channels, and paralinguistic factors such as the emotional state of the speaker. They also aim to be as compact as possible.

Representations used in current speech recognizers (Figure 1.3), concentrate primarily on properties of the speech signal attributable to the shape of the vocal tract rather than to the excitation, whether generated by a vocal-tract constriction or by the larynx. Representations are sensitive to whether the vocal folds are vibrating or not (the voiced/unvoiced distinction), but try to ignore effects due to variations in their frequency of vibration (F_0).

Representations are almost always derived from the short-term power spectrum; that is, the short-term phase structure is ignored. This is primarily because our ears are largely insensitive to phase effects. Consequently, speech communication and recording equipment often does not preserve the phase structure of the original waveform, and such equipment as well as factors such as room acoustics can alter the phase spectrum in ways that would disturb a phase-sensitive speech recognizer even though a human listener would not notice them.

The power spectrum is, moreover, almost always represented on a log scale. When the gain applied to a signal varies, the shape of the log power spectrum is preserved; the

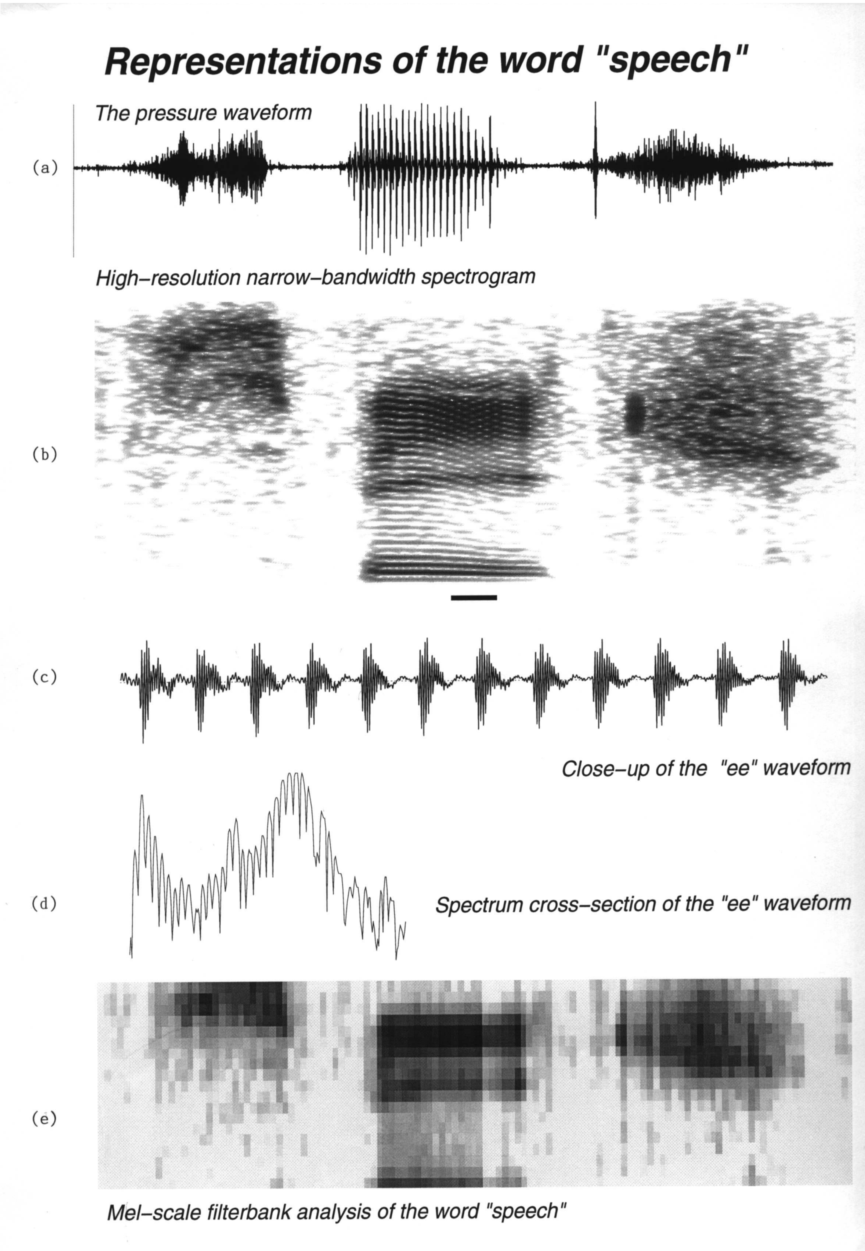

Figure 1.3: Examples of representations used in current speech recognizers. (a) time varying waveform of the word *speech*, showing changes in amplitude (y axis) over time (x axis); (b) speech spectrogram of (a), in terms of frequency (y axis), time (x axis) and amplitude (darkness of the pattern); (c) expanded waveform of the vowel *ee* (underlined in b); (d) spectrum of the vowel *ee*, in terms of amplitude (y axis) and frequency (x axis); and (e) Mel-scale spectrogram.

spectrum is simply shifted up or down. More complicated linear filtering caused, for example, by room acoustics or by variations between telephone lines, which appear as convolutional effects on the waveform and as multiplicative effects on the *linear* power spectrum, become simply additive constants on the log power spectrum. Indeed, a voiced speech waveform amounts to the convolution of a quasi-periodic excitation signal and a time-varying filter determined largely by the configuration of the vocal tract. These two components are easier to separate in the log-power domain, where they are additive. Finally, the statistical distributions of log power spectra for speech have properties convenient for statistically based speech recognition that are not shared by linear power spectra, for example. Because the log of zero is infinite, there is a problem in representing very low energy parts of the spectrum. The log function therefore needs a lower bound both to limit the numerical range and to prevent excessive sensitivity to the low-energy, noise-dominated parts of the spectrum.

Before computing short-term power spectra, the waveform is usually processed by a simple *pre-emphasis* filter giving a 6 dB/octave increase in gain over most of its range to make the average speech spectrum roughly flat.

The short-term spectra are often derived by taking successive overlapping portions of the preemphasized waveform, typically 25 msec long, tapering both ends with a bell-shaped window function, and applying a Fourier transform. The resulting power spectrum has undesirable harmonic fine structure at multiples of F_0. This can be reduced by grouping neighboring sets of components together to form about 20 frequency bands before converting to log power. These bands are often made successively broader with increasing frequency above 1 kHz, usually according to the *technical mel* frequency scale (Davis & Mermelstein, 1980), reflecting the frequency resolution of the human ear. A less common alternative to the process just described is to compute the energy in the bands directly using a bank of digital filters. The results are similar.

Since the shape of the spectrum imposed by the vocal tract is smooth, energy levels in adjacent bands tend to be correlated. Removing the correlation allows the number of parameters to be reduced while preserving the useful information. It also makes it easier to compute reasonably accurate probability estimates in a subsequent statistical matching process. The *cosine transform* (a version of the Fourier transform using only cosine basis functions) converts the set of log energies to a set of *cepstral coefficients*, which turn out to be largely uncorrelated. Compared with the number of bands, typically only about half as many of these cepstral coefficients need be kept. The first cepstral coefficient (C_0) describe the shape of the log spectrum independent of its overall level: C_1 measures the balance between the upper and lower halves of the spectrum, and the higher order coefficients are concerned with increasingly finer features in the spectrum.

To the extent that the vocal tract can be regarded as a lossless unbranched acoustic tube with plane-wave sound propagation along it, its effect on the excitation signal is that of a series of resonances; that is, the vocal tract can be modeled as an *all-pole* filter. For many speech sounds in favorable acoustic conditions, this is a good approximation. A technique known as linear predictive coding (LPC) (Markel & Gray, 1976) or *autoregressive modeling* in effect fits the parameters of an all-pole filter to the speech spectrum, though the spectrum itself need never be computed explicitly. This provides a popular alternative method of deriving cepstral coefficients.

LPC has problems with certain signal degradations and is not so convenient for producing mel-scale cepstral coefficients. Perceptual Linear Prediction (PLP) combines the LPC and filter-bank approaches by fitting an all-pole model to the set of energies (or, strictly, loudness levels) produced by a perceptually motivated filter bank, and then computing the cepstrum from the model parameters (Hermansky, 1990).

Many systems augment information on the short-term power spectrum with information on its rate of change over time. The simplest way to obtain this dynamic information would be to take the difference between consecutive frames. However, this turns out to be too sensitive to random interframe variations. Consequently, linear trends are estimated over sequences of typically five or seven frames (Furui, 1986b).

Some systems go further and estimate acceleration features as well as linear rates of change. These second-order dynamic features need even longer sequences of frames for reliable estimation (Applebaum & Hanson, 1989).

Steady factors affecting the shape or overall level of the spectrum (such as the characteristics of a particular telephone link) appear as constant offsets in the log spectrum and cepstrum. (In a technique called *blind deconvolution* (Stockham, Connon, et al., 1975), cepstrum is computed and this average is substracted from the individual frames.) This method is largely confined to non-real-time experimental systems. Since they are based on differences, however, dynamic features are intrinsically immune to such constant effects. Consequently, while C_0 is usually cast aside, its dynamic equivalent, δC_0, depending only on relative rather than absolute energy levels, is widely used.

If first-order dynamic parameters are passed through a *leaky* integrator, something close to the original static parameters are recovered except that constant and very slowly varying features are reduced to zero, thus giving independence from constant or slowly varying channel characteristics. This technique, amounting to band-pass filtering of sequences of log power spectra and sometimes called *RASTA*, is better suited than blind deconvolution to real-time systems (Hermansky, Morgan, et al., 1993). A similar technique applied to sequences of power spectra before logs are taken is capable of reducing the effect of steady or slowly varying additive noise (Hirsch, Meyer, et al., 1991).

Because cepstral coefficients are largely uncorrelated, a computationally efficient method of obtaining reasonably good probability estimates in the subsequent matching process consists of calculating Euclidean distances from reference model vectors after suitably weighting the coefficients. Various weighting schemes have been used. An empirical scheme that works well derives the weights for the first 16 coefficients from the positive half cycle of a sine wave (Juang, Rabiner, et al., 1986). For PLP cepstral coefficients, weighting each coefficient by its index (root power sum (RPS) weighting) giving C_0 a weight of zero, etc., has proved effective. Statistically based methods weight coefficients by the inverse of their standard deviations computed about their overall means, or preferably computed about the means for the corresponding speech sound and then averaged over all speech sounds (so-called *grand-variance weighting*) (Lippmann, Martin, et al., 1987).

While cepstral coefficients are substantially uncorrelated, a technique called principal components analysis (PCA) can provide a transformation that can completely remove linear dependencies between sets of variables. This method can be used to de-correlate not just sets of energy levels across a spectrum but also combinations of parameter sets such as dynamic and static features, PLP and non-PLP parameters. A double application of PCA with a weighting operation, known as linear discriminant analysis (LDA), can take into account the discriminative information needed to distinguish between speech sounds to generate a set of parameters, sometimes called IMELDA coefficients, suitably weighted for Euclidean-distance calculations. Good performance has been reported with a much reduced set of IMELDA coefficients, and there is evidence that incorporating degraded signals in the analysis can improve robustness to the degradations while not harming performance on undegraded data (Hunt & Lefèbvre, 1989).

Future Directions

The vast majority of major commercial and experimental systems use representations akin to those described here. However, in striving to develop better representations, wavelet transforms (Daubechies, 1990) are being explored, and neural network methods are being used to provide non-linear operations on log spectral representations. Work continues on representations more closely reflecting auditory properties (Greenberg, 1988) and on representations reconstructing articulatory gestures from the speech signal (Schroeter & Sondhi, 1994). This latter work is challenging because there is a one-to-many mapping between the speech spectrum and the articulatory settings that could produce it. It is attractive because it holds out the promise of a small set of smoothly varying parameters that could deal in a simple and principled way with the interactions that occur between neighboring phonemes and with the effects of differences in speaking rate and of carefulness of enunciation.

As we noted earlier, current representations concentrate on the spectrum envelope and ignore fundamental frequency; yet we know that even in isolated-word recognition fundamental frequency contours are an important cue to lexical identity not only in tonal languages such as Chinese but also in languages such as English where they correlate with lexical stress. In continuous speech recognition fundamental frequency contours can potentially contribute valuable information on syntactic structure and on the intentions of the speaker (e.g., *No, I said 2* **5** *7*). The challenges here lie not in deriving fundamental frequency but in knowing how to separate out the various kinds of information that it encodes (speaker identity, speaker state, syntactic structure, lexical stress, speaker intention, etc.) and how to integrate this information into decisions otherwise based on identifying sequences of phonetic events.

The ultimate challenge is to match the superior performance of human listeners over automatic recognizers. This superiority is especially marked when there is little material to allow adaptation to the voice of the current speaker, and when the acoustic conditions are difficult. The fact that it persists even when nonsense words are used shows that it exists at least partly at the acoustic/phonetic level and cannot be explained purely by superior language modeling in the brain. It confirms that there is still much to be done in developing better representations of the speech signal. For additional references, see Rabiner and Schafer (1978) and Hunt (1993).

1.4 Robust Speech Recognition

Richard M. Stern

Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

Robustness in speech recognition refers to the need to maintain good recognition accuracy even when the quality of the input speech is degraded, or when the acoustical, articulatory, or phonetic characteristics of speech in the training and testing environments differ. Obstacles to robust recognition include acoustical degradations produced by additive noise, the effects of linear filtering, nonlinearities in transduction or transmission, as well as impulsive interfering sources, and diminished accuracy caused by changes in articulation produced by the presence of high-intensity noise sources. Some of these sources of variability are illustrated in Figure 1.4. Speaker-to-speaker differences impose a different type of variability, producing variations in speech rate, co-articulation, context, and dialect. Even systems that are designed to be speaker independent exhibit dramatic degradations in recognition accuracy when training and testing conditions differ (Cole, Hirschman, et al., 1992; Juang, 1991).

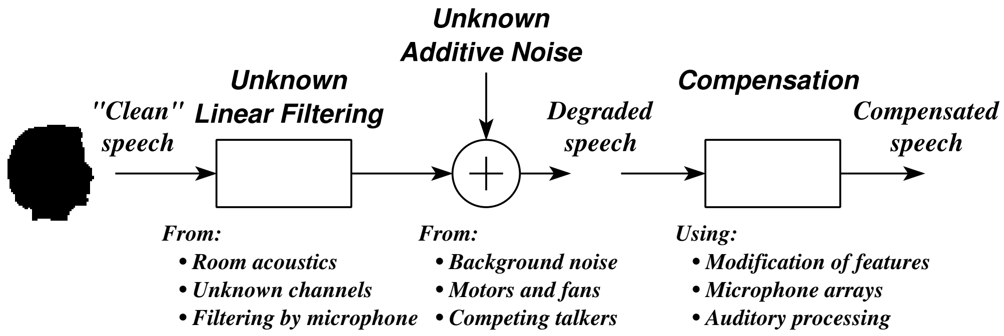

Figure 1.4: Schematic representation of some of the sources of variability that can degrade speech recognition accuracy, along with compensation procedures that improve environmental robustness.

Speech recognition systems have become much more robust in recent years with respect to both speaker variability and acoustical variability. In addition to achieving speaker independence, many current systems can also automatically compensate for modest amounts of acoustical degradation caused by the effects of unknown noise and unknown linear filtering.

As speech recognition and spoken language technologies are being transferred to real applications, the need for greater robustness in recognition technology is becoming increasingly apparent. Nevertheless, the performance of even the best state-of-the art systems tends to deteriorate when speech is transmitted over telephone lines, when the signal-to-noise ratio (SNR) is extremely low (particularly when the unwanted noise consists of speech from other talkers), and when the speaker's native language is not the one with which the system was trained.

Substantial progress has also been made over the last decade in the dynamic adaptation of speech recognition systems to new speakers, with techniques that modify or warp the systems' phonetic representations to reflect the acoustical characteristics of individual speakers (Gauvain & Lee, 1991; Huang & Lee, 1993; Schwartz, Chow, et al., 1987). Speech recognition systems have also become more robust in recent years, particularly with regard to slowly-varying acoustical sources of degradation.

In this section we focus on approaches to environmental robustness. We begin with a discussion of dynamic adaptation techniques for unknown acoustical environments and speakers. We then discuss two popular alternative approaches to robustness, the use of multiple microphones, and the use of signal processing based on models of auditory physiology and perception.

1.4.1 Dynamic Parameter Adaptation

Dynamic adaptation of either the features that are input to the recognition system, or of the system's internally stored representations of possible utterances, is the most direct approach to environmental and speaker adaptation. We discuss separately three different approaches to speaker and environmental adaptation: (1) the use of optimal estimation procedures to obtain new parameter values in the testing conditions; (2) the development of compensation procedures based on empirical comparisons of speech in the training and testing environments; and (3) the use of high-pass filtering of parameter values to improve robustness.

Optimal Parameter Estimation: Many successful robustness techniques are based on a formal statistical model that characterizes the differences between speech used to training and test the system. Parameter values of these models are estimated from samples of speech in the testing environments, and either the features of the incoming speech or the internally-stored representations of speech in the system are modified. Typical structural models for adaptation to acoustical variability assume that speech is corrupted either by additive noise with an unknown power spectrum (Porter & Boll, 1984; Ephraim, 1992; Erell & Weintraub, 1990; Gales & Young, 1992; Lockwood, Boudy, et al., 1992; Bellegarda, de Souza, et al., 1992), or by a combination of additive noise and linear filtering (Acero & Stern, 1990). Much of the early work in robust recognition involved a re-implementation of techniques developed to remove additive noise for the purpose of speech enhancement, as reviewed in section 10.3. The fact that such approaches were able to substantially reduce error rates in machine recognition of speech even though they were largely ineffective in improving human speech intelligibility (when measured objectively) (Lim & Oppenheim, 1979) is one indication of the limited capabilities of automatic speech recognition systems, compared to human speech perception.

Approaches to speaker adaptation are similar in principle, except that the models are more commonly general statistical models of feature variability (Gauvain & Lee, 1991; Huang & Lee, 1993), rather than models of the sources of speaker-to-speaker variability. Solution of the estimation problems frequently requires either analytical or numerical approximations, or the use of iterative estimation techniques such as the estimate-maximize (EM) algorithm (Dempster, Laird, et al., 1977). These approaches have all been successful in applications where the assumptions of the models are reasonably valid, but they are limited in some cases by computational complexity.

Another popular approach is to use knowledge of background noise drawn from examples to transform the means and variances of phonetic models that had been developed for *clean* speech to enable these models to characterize speech in background noise (Varga & Moore, 1990; Gales & Young, 1992). The technique known as parallel model combination(Gales & Young, 1992) extends this approach, providing an analytical model of the degradation that accounts for both additive and convolutional noise. These methods work reasonably well, but they are computationally costly at present, and they rely on accurate estimates of the background noise.

Empirical Feature Comparison: Empirical comparisons of features derived from high-quality speech with features of speech that is simultaneously recorded under degraded conditions can be used (instead of a structural model) to compensate for mismatches between training and testing conditions. In these algorithms, the combined effects of environmental and speaker variability are typically characterized as additive perturbations to the features. Several successful empirically-based robustness algorithms have been described that either apply additive correction vectors to the features derived from incoming speech waveforms (Neumeyer & Weintraub, 1994; Liu, Stern, et al., 1994) or that apply additive correction vectors to the statistical parameters characterizing the internal representations of these features in the recognition system e.g., Anastasakos, Makhoul, et al. (1994); Liu, Stern, et al. (1994). (In the latter case the variances of the templates may also be modified.) Recognition accuracy can be substantially improved by allowing the correction vectors to depend on SNR, specific location in parameter space within a given SNR, or presumed phoneme identity (Neumeyer & Weintraub, 1994; Liu, Stern, et al., 1994). For example, the numerical difference between cepstral coefficients derived on a frame-by-frame basis from high-quality speech and simultaneously recorded speech that is degraded by both noise and filtering primarily reflects the degradations introduced by the filtering at high SNRs, and the effects of the noise at low SNRs. This general approach can be extended to cases when the testing environment is unknown *a priori*, by developing ensembles of correction vectors in parallel for a number of different testing conditions, and by subsequently applying the set of correction vectors (or acoustic models) from the condition that is deemed to be most likely to have produced the incoming speech. In cases where the test condition is not one of the ones used to train correction vectors, recognition accuracy can be further improved by interpolating the correction vectors or statistics representing the best candidate conditions.

Empirically-derived compensation procedures are extremely simple, and they are quite effective in cases when the testing conditions are reasonably similar to one of the conditions used to develop correction vectors. For example, in a recent evaluation using speech from a number of unknown microphones in a 5000-word continuous dictation task, the use of adaptation techniques based on empirical comparisons of feature values reduced the error rate by 40% relative to a baseline system with only cepstral mean normalization (described below). Nevertheless, the empirical approaches have the disadvantage of requiring *stereo* databases of speech that are simultaneously recorded in the training environment and the testing environment.

Cepstral High-pass Filtering: The third major adaptation technique is cepstral high-pass filtering, which provides a remarkable amount of robustness at almost zero computational cost (Hermansky, Morgan, et al., 1991; Hirsch, Meyer, et al., 1991). In the well-known RASTA method (Hermansky, Morgan, et al., 1991), a high-pass (or band-pass) filter is applied to a log-spectral representation of speech such as the cepstral coefficients. In cepstral mean normalization (CMN), high-pass filtering is accomplished by subtracting the short-term average of cepstral vectors from the incoming cepstral coefficients.

The original motivation for the RASTA and CMN algorithms is discussed in section 1.3. These algorithms compensate directly for the effects of unknown linear filtering because they force the average values of cepstral coefficients to be zero in both the training and testing domains, and hence equal to each other. An extension to the RASTA algorithm known as J-RASTA (Koehler, Morgan, et al., 1994) can also compensate for noise at low SNRs. In an evaluation using 13 isolated digits over telephone lines, it was shown (Koehler, Morgan, et al., 1994) that the J-RASTA method reduced error rates by as much as 55 percent relative to RASTA when both noise and filtering effects are present. Cepstral high-pass filtering is so inexpensive and effective that it is currently embedded in some form in virtually all systems that are required to perform robust recognition.

1.4.2 Use of Multiple Microphones

Further improvements in recognition accuracy can be obtained at lower SNRs by the use of multiple microphones. As noted in the discussion on speech enhancement in section 10.3, microphone arrays can, in principle, produce directionally sensitive gain patterns that can be adjusted to increase sensitivity to the speaker and reduce sensitivity in the direction of competing sound sources. In fact, results of recent pilot experiments in office environments (Che, Lin, et al., 1994; Sullivan & Stern, 1993) confirm that the use of delay-and-sum beamformers in combination with a post-processing algorithm that compensates for the spectral coloration introduced by the array itself can reduce recognition error rates by as much as 61%.

Array processors that make use of the more general minimum mean square error (MMSE)-based classical adaptive filtering techniques can work well when signal degradation is dominated by additive independent noise, but they do not perform well in reverberant environments when the distortion is at least in part a delayed version of the desired speech signal (Peterson, 1989; Alvarado & Silverman, 1990). (This problem can be avoided by adapting only during non-speech segments: Van Compernolle, 1990.)

A third approach to microphone array processing is the use of cross-correlation-based algorithms, which have the ability to reinforce the components of a sound field arriving from a particular azimuth angle. These algorithms are appealing because they are similar to the processing performed by the human binaural system, but thus far they have demonstrated only a modest superiority over the simpler delay-and-sum approaches (Sullivan & Stern, 1993).

1.4.3 Use of Physiologically Motivated Signal Processing

A number of signal processing schemes have been developed for speech recognition systems that mimic various aspects of human auditory physiology and perception (e.g., Cohen, 1989; Ghitza, 1988; Lyon, 1982; Seneff, 1988; Hermansky, 1990; Patterson, Robinson, et al., 1991). Such *auditory models* typically consist of a bank of bandpass filters (representing auditory frequency selectivity) followed by nonlinear interactions within and across channels (representing hair-cell transduction, lateral suppression, and other effects). The nonlinear processing is (in some cases) followed by a mechanism to extract detailed timing information as a function of frequency (Seneff, 1988; Duda, Lyon, et al., 1990).

Recent evaluations indicate that auditory models can indeed provide better recognition accuracy than traditional cepstral representations when the quality of the incoming speech degrades, or when training and testing conditions differ (Hunt & Lefèbvre, 1989; Meng & Zue, 1990). Nevertheless, auditory models have not yet been able to demonstrate better recognition accuracy than the most effective dynamic adaptation algorithms, and conventional adaptation techniques are far less computationally costly (Ohshima, 1993). It is possible that the success of auditory models has been limited thus far because most of the evaluations were performed using hidden Markov model classifiers, which are not well matched to the statistical properties of features produced by auditory models. Other researchers suggest that we have not yet identified the features of the models' outputs that will ultimately provide superior performance. The approach of auditory modeling continues to merit further attention, particularly with the goal of resolving these issues.

1.4.4 Future Directions

Despite its importance, robust speech recognition has become a vital area of research only recently. To date, major successes in environmental adaptation have been limited either to relatively benign domains (typically with limited amounts of quasi-stationary additive noise and/or linear filtering, or to domains in which a great deal of environment-specific training data are available). Speaker adaptation algorithms have been successful in providing improved recognition for native speakers languages other than the one with which a system is trained, but recognition accuracy obtained using non-native speakers remains substantially worse, even with speaker adaptation, (e.g., Pallett, Fiscus, et al. (1995)).

At present, it is fair to say that hardly any of the major limitations to robust recognition cited in section 1.1 have been satisfactorily resolved. It is suggested that success in the following key problem areas is likely to accelerate the development and deployment of practical speech-based applications.

Speech over Telephone Lines: Recognition of telephone speech is difficult because each telephone channel has its own unique SNR and frequency response. Speech over telephone lines can be further corrupted by transient interference and nonlinear distortion. Telephone-based applications must be able to adapt to new channels on the basis of a very small amount of channel-specific data.

Low-SNR Environments: Even with state-of-the art compensation techniques, recognition accuracy degrades when the channel SNR decreases below about 15 dB, even though humans can obtain excellent recognition accuracy at lower SNRs.

Co-channel Speech Interference: Interference by other talkers poses a much more difficult challenge to robust recognition than interference by broadband noise sources. So far, efforts to exploit speech-specific information to reduce the effects of co-channel interference by other talkers have been largely unsuccessful.

Rapid Adaptation for Non-native Speakers: In today's pluralistic and highly mobile society, successful spoken-language applications must be able to cope with the speech of non-native as well as native speakers. Continued development of non-intrusive rapid adaptation to the accents of non-native speakers will be needed to ensure commercial success.

Common Speech Corpora with Realistic Degradations: Continued rapid progress in robust recognition will depend on the formulation, collection, transcription, and dissemination of speech corpora that contain realistic examples of the degradations encountered in practical environments. The selection of appropriate tasks and domains for shared database resources is best accomplished through the collaboration of technology developers, applications developers, and end users. The contents of these databases should be realistic enough to be useful as an impetus for solutions to actual problems, even in cases for which it may be difficult to *calibrate* the degradation for the purpose of evaluation.

1.5 HMM Methods in Speech Recognition

Renato De Mori[a] & Fabio Brugnara[b]

[a] McGill University, Montréal, Québéc, Canada
[b] Istituto per la Ricerca Scientifica e Tecnologica, Trento, Italy

Modern architectures for Automatic Speech Recognition (ASR) are mostly software architectures generating a sequence of word hypotheses from an acoustic signal. The most popular algorithms implemented in these architectures are based on statistical methods. Other approaches can be found in Waibel and Lee (1990) where a collection of papers describes a variety of systems with historical reviews and mathematical foundations.

A vector y_t of acoustic features is computed every 10 to 30 msec. Details of this component can be found in section 1.3. Various possible choices of vectors together with their impact on recognition performance are discussed in Haeb-Umbach, Geller, et al. (1993).

Sequences of vectors of acoustic parameters are treated as observations of acoustic word models used to compute $p(\boldsymbol{y}_1^T|W)$,[2] the probability of observing a sequence $\boldsymbol{y}_1^T$ of vectors when a word sequence W is pronounced. Given a sequence $\boldsymbol{y}_1^T$, a word sequence $\widehat{W}$ is generated by the ASR system with a search process based on the rule:

$$\widehat{W} = \arg\max_{W} \; p(\boldsymbol{y}_1^T|W)\, p(W)$$

$\widehat{W}$ corresponds to the candidate having maximum a-posteriori probability (MAP). $p(\boldsymbol{y}_1^T|W)$ is computed by Acoustic Models (AM), while $p(W)$ is computed by Language Models (LM).

For large vocabularies, search is performed in two steps. The first generates a word lattice of the *n-best* word sequences with simple models to compute approximate likelihoods in real-time. In the second step more accurate likelihoods are compared with a limited number of hypotheses. Some systems generate a single word sequence hypothesis with a single step. The search produces an hypothesized word sequence if the task is dictation. If the task is understanding then a conceptual structure is obtained with a process that may involve more than two steps. Ways for automatically learning and extracting these structures are described in Kuhn, De Mori, et al. (1994).

[2] Here and in the following, the notation $\boldsymbol{y}_h^k$ stands for the sequence $[y_h, y_{h+1}, \ldots, y_k]$.

1.5.1 Acoustic Models

In a statistical framework, an inventory of elementary probabilistic models of basic linguistic units (e.g., phonemes) is used to build word representations. A sequence of acoustic parameters, extracted from a spoken utterance, is seen as a realization of a concatenation of elementary processes described by hidden Markov models (HMMs). An HMM is a composition of two stochastic processes, a *hidden* Markov chain, which accounts for *temporal* variability, and an observable process, which accounts for *spectral* variability. This combination has proven to be powerful enough to cope with the most important sources of speech ambiguity, and flexible enough to allow the realization of recognition systems with dictionaries of tens of thousands of words.

Structure of a Hidden Markov Model

A hidden Markov model is defined as a pair of stochastic processes $(\boldsymbol{X}, \boldsymbol{Y})$. The $\boldsymbol{X}$ process is a first order Markov chain, and is not directly observable, while the $\boldsymbol{Y}$ process is a sequence of random variables taking values in the space of acoustic parameters, or *observations*.

Two formal assumptions characterize HMMs as used in speech recognition. The *first-order Markov hypothesis* states that history has no influence on the chain's future evolution if the present is specified, and the *output independence hypothesis* states that neither chain evolution nor past observations influence the present observation if the last chain transition is specified.

Letting $y \in \mathcal{Y}$ be a variable representing observations and $i, j \in \mathcal{X}$ be variables representing model states, the model can be represented by the following parameters:

$$\begin{array}{lcll} A & \equiv & \{a_{i,j} | i, j \in \mathcal{X}\} & \text{transition probabilities} \\ B & \equiv & \{b_{i,j} | i, j \in \mathcal{X}\} & \text{output distributions} \\ \Pi & \equiv & \{\pi_i | i \in \mathcal{X}\} & \text{initial probabilities} \end{array}$$

with the following definitions:

$$\begin{array}{rcl} a_{i,j} & \equiv & p\left(X_t = j | X_{t-1} = i\right) \\ b_{i,j}(y) & \equiv & p\left(Y_t = y | X_{t-1} = i, X_t = j\right) \\ \pi_i & \equiv & p\left(X_0 = i\right) \end{array}$$

A useful tutorial on the topic can be found in Rabiner (1989).

Types of Hidden Markov Models

HMMs can be classified according to the nature of the elements of the B matrix, which are distribution functions.

Distributions are defined on finite spaces in the so called *discrete HMMs*. In this case, observations are vectors of symbols in a finite alphabet of N different elements. For each one of the Q vector components, a discrete density $\{w(k)|k = 1, \ldots, N\}$ is defined, and the distribution is obtained by multiplying the probabilities of each component. Notice that this definition assumes that the different components are independent. Figure 1.5 shows an example of a discrete HMM with one-dimensional observations. Distributions are associated with model transitions.

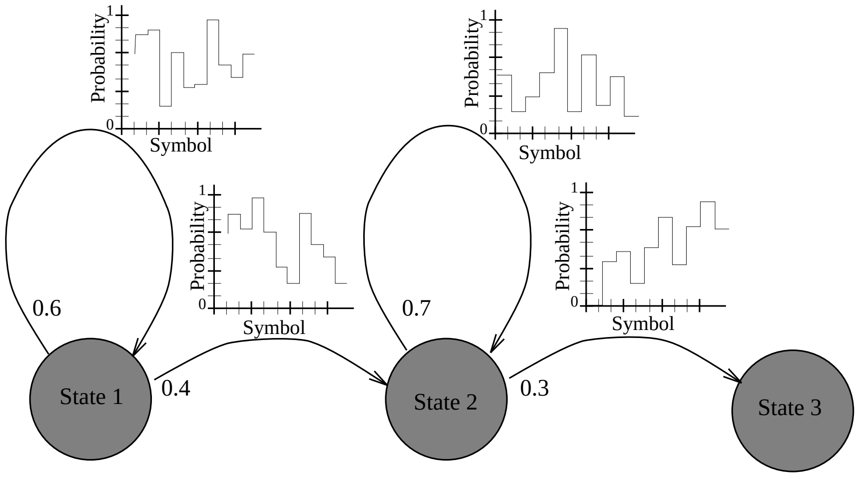

Figure 1.5: Example of a discrete HMM. A transition probability and an output distribution on the symbol set is associated with every transition.

Another possibility is to define distributions as probability densities on continuous observation spaces. In this case, strong restrictions have to be imposed on the functional form of the distributions, in order to have a manageable number of statistical parameters to estimate. The most popular approach is to characterize the model transitions with mixtures of base densities g of a family G having a simple parametric form. The base densities $g \in G$ are usually Gaussian or Laplacian, and can be parameterized by the mean vector and the covariance matrix. HMMs with these kinds of distributions are usually referred to as *continuous HMMs*. In order to model complex distributions in this way a rather large number of base densities has to be used in every mixture. This may require a very large training corpus of data for the estimation of the distribution parameters. Problems arising when the available corpus is not large enough can be alleviated by sharing distributions among transitions of different models. In *semicontinuous HMMs* Huang, Ariki, et al. (1990), for example, all mixtures are expressed in terms of a common set of base densities. Different mixtures are characterized only by different weights.

A common generalization of semicontinuous modeling consists of interpreting the input vector y as composed of several components $y[1], \ldots, y[Q]$, each of which is associated with a different set of base distributions. The components are assumed to be statistically independent, hence the distributions associated with model transitions are products of the component density functions.

Computation of probabilities with discrete models is faster than with continuous models, nevertheless it is possible to speed up the mixture densities computation by applying vector quantization (VQ) on the gaussians of the mixtures (Bocchieri, 1993).

Parameters of statistical models are estimated by iterative learning algorithms (Rabiner, 1989) in which the likelihood of a set of training data is guaranteed to increase at each step.

Bengio, DeMori, et al. (1992) propose a method for extracting additional acoustic parameters and performing transformations of all the extracted parameters using a Neural Network (NN) architecture whose weights are obtained by an algorithm that, at the same time, estimates the coefficients of the distributions of the acoustic models. Estimation is driven by an optimization criterion that tries to minimize the overall recognition error.

1.5.2 Word and Unit Models

Words are usually represented by networks of phonemes. Each path in a word network represents a pronunciation of the word.

The same phoneme can have different acoustic distributions of observations if pronounced in different contexts. *Allophone* models of a phoneme are models of that phoneme in different contexts. The decision as to how many allophones should be considered for a given phoneme may depend on many factors, e.g., the availability of enough training data to infer the model parameters.

A conceptually interesting approach is that of *polyphones* (Shukat-Talamazzini, Niemann, et al., 1992). In principle, an allophone should be considered for every different word in which a phoneme appears. If the vocabulary is large, it is unlikely that there are enough data to train all these allophone models, so models for allophones of phonemes are considered at a different level of detail (word, syllable, triphone, diphone, context independent phoneme). Probability distributions for an allophone having a certain degree of generality can be obtained by mixing the distributions of more detailed allophone models. The loss in specificity is compensated by a more robust estimation of the statistical parameters due to the increasing of the ratio between training data and free parameters to estimate.

Another approach consists of choosing allophones by *clustering* possible contexts. This choice can be made automatically with Classification and Regression Trees (CART). A CART is a binary tree having a phoneme at the root and, associated with each node n_i, a question Q_i about the context. Questions Q_i are of the type, "Is the previous phoneme a nasal consonant?" For each possible answer (*YES* or *NO*) there is a link to another node with which other questions are associated. There are algorithms for growing and pruning CARTs based on automatically assigning questions to a node from a manually determined pool of questions. The leaves of the tree may be simply labeled by an allophone symbol. Papers by Bahl, de Souza, et al. (1991) and Hon and Lee (1991) provide examples of the application of this concept and references to the description of a formalism for training and using CARTs.

Each allophone model is an HMM made of states, transitions and probability distributions. In order to improve the estimation of the statistical parameters of these models, some distributions can be the same or tied. For example, the distributions for the central portion of the allophones of a given phoneme can be tied reflecting the fact that they represent the stable (context-independent) physical realization of the central part of the phoneme, uttered with a stationary configuration of the vocal tract.

In general, all the models can be built by sharing distributions taken from a pool of, say, a few thousand cluster distributions called *senones*. Details on this approach can be found in Hwang and Huang (1993).

Word models or allophone models can also be built by concatenation of basic structures made by states, transitions and distributions. These units, called *fenones*, were introduced by Bahl, Brown, et al. (1993). Richer models of the same type but using more sophisticated building blocks, called *multones*, are described in Bahl, Bellegarda, et al. (1993).

Another approach consists of having clusters of distributions characterized by the same set of Gaussian probability density functions. Allophone distributions are built by considering mixtures with the same components but with different weights (Digalakis & Murveit, 1994).

1.5.3 Language Models

The probability $p(W)$ of a sequence of words $W = w_1, \ldots, w_L$ is computed by a Language Model (LM). In general $p(W)$ can be expressed as follows:

$$p(W) = p(w_1, .., w_n) = \prod_{i=1}^{n} p(w_i | w_0, .., w_{i-1})$$

Motivations for this approach and methods for computing these probabilities are described in the following section.

1.5.4 Generation of Word Hypotheses

Generation of word hypotheses can result in a single sequence of words, in a collection of the *n-best* word sequences, or in a lattice of partially overlapping word hypotheses.

This generation is a search process in which a sequence of vectors of acoustic features is compared with word models. In this section, some distinctive characteristics of the computations involved in speech recognition algorithms will be described, first focusing on the case of a single-word utterance, and then considering the extension to continuous speech recognition.

In general, the speech signal and its transformations do not exhibit clear indication of word boundaries, so word boundary detection is part of the hypothesization process carried out as a search. In this process, all the word models are compared with a sequence of acoustic features. In the probabilistic framework, "comparison" between an acoustic sequence and a model involves the computation of the probability that the model assigns to the given sequence. This is the key ingredient of the recognition process. In this computation, the following quantities are used:

$\alpha_t(\boldsymbol{y}_1^T, i)$: probability of having observed the partial sequence $\boldsymbol{y}_1^t$ and being in state i at time t

$$\alpha_t(\boldsymbol{y}_1^T, i) \equiv \begin{cases} p\,(X_0 = i)\,, & t = 0 \\ p\,(X_t = i, \boldsymbol{Y}_1^t = \boldsymbol{y}_1^t)\,, & t > 0 \end{cases}$$

$\beta_t(\boldsymbol{y}_1^T, i)$: probability of observing the partial sequence $\boldsymbol{y}_{t+1}^T$ given that the model is in state i at time t

$$\beta_t(\boldsymbol{y}_1^T, i) \equiv \begin{cases} p\left(\boldsymbol{Y}_{t+1}^T = \boldsymbol{y}_{t+1}^T | X_t = i\right), & t < T \\ 1, & t = T \end{cases}$$

$\psi_t(\boldsymbol{y}_1^T, i)$: probability of having observed the partial sequence $\boldsymbol{y}_1^t$ along the best path ending in state i at time t:

$$\psi_t(\boldsymbol{y}_1^T, i) \equiv \begin{cases} p\left(X_0 = i\right), & t = 0 \\ \max\limits_{\boldsymbol{i}_0^{t-1}} p\left(\boldsymbol{X}_0^{t-1} = \boldsymbol{i}_0^{t-1}, X_t = i, \boldsymbol{Y}_1^t = \boldsymbol{y}_1^t\right) & t > 0 \end{cases}$$

α and β can be used to compute the total emission probability $p(\boldsymbol{y}_1^T|W)$ as

$$p(\boldsymbol{Y}_1^T = \boldsymbol{y}_1^T) = \sum_i \alpha_T(\boldsymbol{y}_1^T, i) \tag{1.1}$$

$$= \sum_i \pi_i \beta_0(\boldsymbol{y}_1^T, i) \tag{1.2}$$

An approximation for computing this probability consists of following only the path of maximum probability. This can be done with the ψ quantity:

$$\Pr{}^*[\boldsymbol{Y}_1^T = \boldsymbol{y}_1^T] = \max_i \psi_T(\boldsymbol{y}_1^T, i) \tag{1.3}$$

The computations of all the above probabilities share a common framework, employing a matrix called a *trellis*, depicted in Figure 1.6. For the sake of simplicity, we can assume that the HMM in Figure 1.6 represents a word and that the input signal corresponds to the pronunciation of an isolated word.

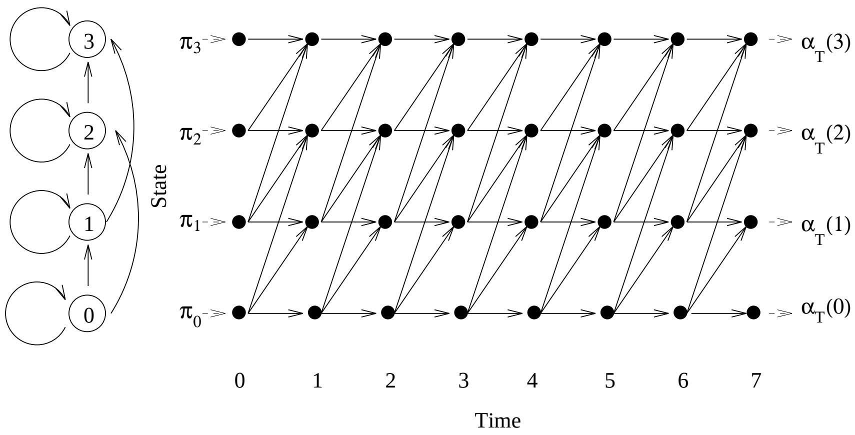

Figure 1.6: A state-time trellis.

Every trellis column holds the values of one of the just introduced probabilities for a partial sequence ending at different time instants, and every interval between two columns corresponds to an input frame. The arrows in the trellis represent model transitions composing possible paths in the model from the initial time instant to the final one. The computation proceeds in a column-wise manner, at every time frame updating the scores of the nodes in a column by means of recursion formulas which involve the values of an adjacent column, the transition probabilities of the models, and the values of the output distributions for the corresponding frame. For α and ψ coefficients, the computation starts from the leftmost column, whose values are initialized with the values of π_i, and ends at the opposite side, computing the final value with (1.1) or (1.3). For the β coefficients, the computation goes from right to left.

The algorithm for computing ψ coefficients is known as the *Viterbi algorithm*, and can be seen as an application of dynamic programming for finding a maximum probability path in a graph with weighted arcs. The recursion formula for its computation is the following:

$$\psi_t(\boldsymbol{y}_1^T, i) = \begin{cases} \pi_i, & t = 0 \\ \max_j \psi_{t-1}(\boldsymbol{y}_1^T, j) a_{j,i} b_{j,i}(y_t), & t > 0 \end{cases}$$

By keeping track of the state j giving the maximum value in the above recursion formula, it is possible, at the end of the input sequence, to retrieve the states visited by the best path, thus performing a sort of time-alignment of input frames with models states.

All these algorithms have a time complexity $O(MT)$, where M is the number of transitions with non-zero probability and T is the length of the input sequence. M can be at most equal to S^2, where S is the number of states in the model, but is usually much lower, since the transition probability matrix is generally sparse. In fact, a common choice in speech recognition is to impose severe constraints on the allowed state sequences, for example $a_{i,j} = 0$ for $j < i, j > i + 2$, as is the case of the model in Figure 1.6.

In general, recognition is based on a search process which takes into account all the possible segmentations of the input sequence into words, and the a-priori probabilities that the LM assigns to sequences of words.

Good results can be obtained with simple LMs based on bigram or trigram probabilities. As an example, let us consider a bigram language model. This model can be conveniently incorporated into a finite state automaton as shown in Figure 1.7, where dashed arcs correspond to transitions between words with probabilities of the LM.

After substitution of the word-labeled arcs with the corresponding HMMs, the resulting automaton becomes a big HMM itself, on which a Viterbi search for the most probable

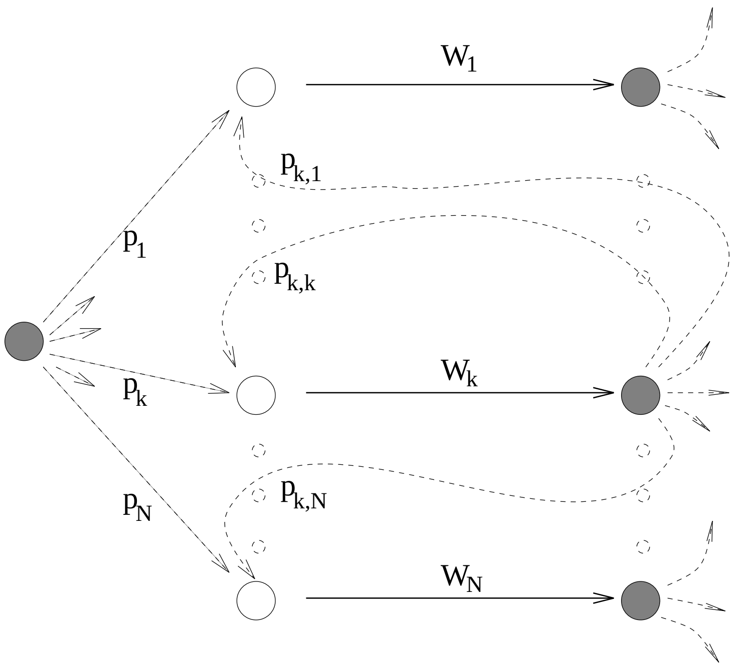

Figure 1.7: Bigram LM represented as a weighted word graph. $p_{h,k}$ stands for $p(W_k|W_h)$, p_h stands for $p(W_h)$. The leftmost node is the starting node, rightmost ones are finals.

path, given an observation sequence, can be carried out. The dashed arcs are to be treated as *empty transitions*, i.e., transitions without an associated output distribution. This requires some generalization of the Viterbi algorithm. During the execution of the Viterbi algorithm, a minimum of backtracking information is kept to allow the reconstruction of the best path in terms of word labels. Note that the solution provided by this search is *suboptimal* in the sense that it gives the probability of a single state sequence of the composite model and not the total emission probability of the best word model sequence. In practice, however, it has been observed that the path probabilities computed with the above mentioned algorithms exhibit a dominance property, consisting of a single state sequence accounting for most of the total probability (Merhav & Ephraim, 1991).

The composite model grows with the vocabulary, and can lead to large search spaces. Nevertheless the uneven distribution of probabilities among different paths can help. It turns out that, when the number of states is large, at every time instant, a large portion of states have an accumulated likelihood which is much less than the highest one, so that it is very unlikely that a path passing through one of these states would become the best path at the end of the utterance. This consideration leads to a complexity reduction technique called *beam search* (Ney, Mergel, et al., 1992), consisting of neglecting states whose accumulated score is lower than the best one minus a given threshold. In this way, computation needed to expand *bad* nodes is avoided. It is clear from the naivety of the pruning criterion that this reduction technique has the undesirable property of being *not admissible*, possibly causing the loss of the best path. In practice, good tuning of the beam threshold results in a gain in speed by an order of magnitude, while introducing a negligible amount of search errors.

When the dictionary is of the order of tens of thousands of words, the network becomes too big, and others methods have to be considered.

At present, different techniques exist for dealing with very large vocabularies. Most of them use multi-pass algorithms. Each pass prepares information for the next one, reducing the size of the search space. Details of these methods can be found in Alleva, Huang, et al. (1993); Aubert, Dugast, et al. (1994); Murveit, Butzberger, et al. (1993); Kubala, Anastasakos, et al. (1994).

In a first phase a set of candidate interpretations is represented in an object called *word lattice*, whose structure varies in different systems: it may contain only hypotheses on the location of words, or it may carry a record of acoustic scores as well. The construction of the word lattice may involve only the execution of a Viterbi beam-search with memorization of word scoring and localization, as in Aubert, Dugast, et al. (1994), or may itself require multiple steps, as in Alleva, Huang, et al. (1993); Murveit, Butzberger, et al. (1993); Kubala, Anastasakos, et al. (1994). Since the word lattice is only an intermediate result, to be inspected by other detailed methods, its generation is performed with a bigram language model, and often with simplified acoustic models.

The word hypotheses in the lattice are scored with a more accurate language model, and sometimes with more detailed acoustic models. Lattice rescoring may require new calculations of HMM probabilities (Murveit, Butzberger, et al., 1993), may proceed on the basis of precomputed probabilities only (Aubert, Dugast, et al., 1994; Alleva, Huang, et al., 1993), or even exploit acoustic models which are not HMMs (Kubala, Anastasakos, et al., 1994). In Alleva, Huang, et al. (1993), the last step is based on an A^* search (Nilsson, 1971) on the word lattice, allowing the application of a *long distance language model*, i.e., a model where the probability of a word may not only depend on its immediate predecessor. In Aubert, Dugast, et al. (1994) a dynamic programming algorithm, using trigram probabilities, is performed.

A method which does not make use of the word lattice is presented in Paul (1994). Inspired by one of the first methods proposed for continuous speech recognition (CSR) (Jelinek, 1969), it combines both powerful language modeling and detailed acoustic modeling in a single step, performing an A^* based search.

1.5.5 Future Directions

Interesting software architectures for ASR have been recently developed. They provide acceptable recognition performance almost in real time for dictation of large vocabularies (more than 10,000 words). Pure software solutions require, at the moment, a considerable amount of central memory. Special boards make it possible to run interesting applications on PCs.

There are aspects of the best current systems that still need improvement. The best systems do not perform equally well with different speakers and different speaking environments. Two important aspects, namely recognition in noise and speaker adaptation, are discussed in section 1.4. They have difficulty in handling out of vocabulary words, hesitations, false starts and other phenomena typical of spontaneous speech. Rudimentary understanding capabilities are available for speech understanding in limited domains. Key research challenges for the future are acoustic robustness, use of better acoustic features and models, use of multiple word pronunciations and efficient constraints for the access of a very large lexicon, sophisticated and multiple language models capable of representing various types of contexts, rich methods for extracting conceptual representations from word hypotheses and automatic learning methods for extracting various types of knowledge from corpora.

1.6 Language Representation

Salim Roukos

IBM T. J. Watson Research Center, Yorktown Heights, New York, USA

A speech recognizer converts the observed acoustic signal into the corresponding orthographic representation of the spoken sentence. The recognizer chooses its guess from a finite vocabulary of *words* that can be recognized. For simplicity, we assume that a word is uniquely identified by its spelling.[3]

Dramatic progress has been demonstrated in solving the speech recognition problem via the use of a statistical model of the joint distribution $p(W, O)$ of the sequence of spoken words W and the corresponding observed sequence of acoustic information O. This approach, pioneered by the IBM Continuous Speech Recognition group, is called the *source-channel model*. In this approach, the speech recognizer determines an estimate $\hat{W}$ of the identity of the spoken word sequence from the observed acoustic evidence O by using the a posteriori distribution $p(W|O)$. To minimize its error rate, the recognizer chooses that word sequence that maximizes the a posteriori distribution:

$$\hat{W} = \arg\max_W \; p(W|O) = \arg\max_W \; \frac{p(W)p(O|W)}{p(O)}$$

where $p(W)$ is the probability of the sequence of n-words W and $p(O|W)$ is the probability of observing the acoustic evidence O when the sequence W is spoken. The a priori distribution $p(W)$ of what words might be spoken (the source) is referred to as a language model (LM). The observation probability model $p(O|W)$ (the channel) is called the acoustic model. We discuss in this section, various approaches and issues for building the language model.

The source-channel model has also been used in optical character recognition (OCR) where the observation sequence is the image of the printed characters, in handwriting recognition where the observation is the sequence of strokes on a tablet, or in machine translation (MT) where the observation is a sequence of words in one language and W represents the desired translation in another language. For all these applications, a language model is key. Therefore, the work on language modeling has a wide spectrum of applications.

[3]For example, we treat as the same word the present and past participle of the verb read (*I read* vs. *I have read*) in the LM while the acoustic model will have different models corresponding to the different pronunciations.

1.6.1 Trigram Language Model

For a given word sequence $W = \{w_1, .., w_n\}$ of n words, we rewrite the LM probability as:

$$p(W) = p(w_1, .., w_n) = \prod_{i=1}^{n} p(w_i | w_0, .., w_{i-1})$$

where w_0 is chosen appropriately to handle the initial condition. The probability of the next word w_i depends on the history h_i of words that have been spoken so far. With this factorization the complexity of the model grows exponentially with the length of the history. To have a more practical and parsimonious model, only some aspects of the history are used to affect the probability of the next word. One way[4] to achieve this is to use a mapping $\phi(\ )$ that divides the space of histories into K equivalence classes. Then we can use as a model:

$$p(w_i | h_i) \approx p(w_i | \phi(h_i)).$$

Some of the most successful models of the past two decades are the simple *n-gram* models, particularly the trigram model ($n = 3$) where only the most recent two words of the history are used to condition the probability of the next word. The probability of a word sequence becomes:

$$p(W) \approx \prod_{i=1}^{n} p(w_i | w_{i-2}, w_{i-1}).$$

To estimate the trigram probabilities, one can use a large corpus of text, called the *training corpus* to estimate trigram frequencies:

$$f_3(w_3 | w_1, w_2) = \frac{c_{123}}{c_{12}}$$

where c_{123} is the number of times the sequence of words $\{w_1, w_2, w_3\}$ is observed and c_{12} is the number of times the sequence $\{w_1, w_2\}$ is observed. For a vocabulary size V there are V^3 possible trigrams, which for 20,000 words translates to 8 trillion trigrams. Many of these trigrams will not be seen in the training corpus. So these unseen trigrams will have zero probability using the trigram frequency as an estimate of the trigram probability. To solve this problem one needs a smooth estimate of the probability of unseen events. This can be done by linear interpolation of trigram, bigram, and unigram frequencies and a uniform distribution on the vocabulary.

[4] Instead of having a single partition of the space of histories, one can use the exponential family to define a set of features that are used for computing the probability of an event. See the discussion on Maximum Entropy in Lau, Rosenfeld, et al. (1993); Darroch and Ratcliff (1972); Berger, Della Pietra, et al. (1994) for more details.

$$p(w_3|w_1, w_2) = \lambda_3 f_3(w_3|w_1, w_2) + \lambda_2 f_2(w_3|w_2) + \lambda_1 f_1(w_3) + \lambda_0 \frac{1}{V}$$

where $f_2(\ )$ and $f_1(\ )$ are estimated by the ratio of the appropriate bigram and unigram counts. The weights of the linear interpolation are estimated by maximizing the probability of new *held-out* data different from the data used to estimate the *n-gram* frequencies. The forward-backward algorithm can be used to perform this maximum likelihood estimation problem.

In general, one uses more than one λ vector; one may want to rely more on the trigram frequencies for those histories that have a high count as compared to those histories that have a low count in the training data. To achieve this, one can use a bucketing scheme on the bigram and unigram counts of the history $b(c_{12}, c_2)$ to determine the interpolation weight vector $\lambda_{b(c_{12},c_2)}$. Typically, 100 to 1,000 buckets are used. This method of smoothing is called *deleted interpolation* (Bahl, Jelinek, et al., 1983). Other smoothing schemes have been proposed such as backing-off, co-occurrence smoothing, and count re-estimation. In the work on language modeling, corpora varying in size from about a million to 500 million words have been used to build trigram models. Vocabulary sizes varying from 1,000 to 267,000 words have also been used. We discuss in the following section the perplexity measure for evaluating a language model.

1.6.2 Perplexity

Given two language models, one needs to compare them. One way is to use them in a recognizer and find the one that leads to the lower recognition error rate. This remains the best way of evaluating a language model. But to avoid this expensive approach one can use the information theory quantity of entropy to get an estimate of how good a LM might be. The basic idea is to average the log probability on per word basis for a piece of new text not used in building the language model.

Denote by p the true distribution, that is unknown to us, of a segment of new text x of k words. Then the entropy on a per word basis is defined

$$H = \lim_{n \to \infty} -\frac{1}{k} \sum_x p(x) \log_2 p(x)$$

If every word in a vocabulary of size $|V|$ is equally likely then the entropy would be $\log_2 |V|$; for other distributions of the words $H \leq \log_2 |V|$.

To determine the probability of this segment of text we will use our language model denoted by $\tilde{p}$ which is different from the true unknown distribution p of the new text. We can compute the average *logprob* on a per word basis defined as:

Domain	Perplexity
Radiology	20
Emergency medicine	60
Journalism	105
General English	247

Table 1.2: Perplexity of trigram models for different domains.

$$lp_k = -\frac{1}{k}\sum_{i=1}^{k} \log_2 \tilde{\mathrm{p}}(w_i|h_i)$$

One can show that $\lim_{k->\inf} lp_k = lp \geq H$; i.e., the average *logprob* is no lower than the entropy of the test text. Obviously our goal is to find that LM which has an average *logprob* that is as close as possible to the entropy of the text.

A related measure to the average *logprob* called *perplexity* is used to evaluate a LM. Perplexity is defined as 2^{lp}. Perplexity is, crudely speaking, a measure of the size of the set of words from which the next word is chosen given that we observe the history of spoken words. The perplexity of a LM depends on the domain of discourse. For radiology reports, one expects less variation in the sentences than in general English. Table 1.2 shows the perplexity of several domains for large vocabulary (20,000 to 30,000 words) dictation systems. The lowest perplexity that has been published on the standard Brown Corpus of 1 million words of American English is about 247 which corresponds to an entropy of 1.75 bits/character.

1.6.3 Vocabulary Size

The error rate of a speech recognizer is no less than the percentage of spoken words that are not in its vocabulary V. So a major part of building a language model is to select a vocabulary that will have maximal coverage on new text spoken to the recognizer. This remains a human intensive effort. A corpus of text is used in conjunction with dictionaries to determine appropriate vocabularies. A tokenizer[5] (a system that

[5]Tokenizing English is fairly straightforward since white space separates words and simple rules can capture many of the punctuations. Special care has to be taken for abbreviations. For oriental languages such as Japanese and Chinese word segmentation is a more complicated problem since space is not used between words.

Vocabulary Size	Static Coverage
20,000	94.1%
64,000	98.7%
100,000	99.3%
200,000	99.4%

Table 1.3: Static coverage of unseen text as a function of vocabulary size.

Number of added words	Text size	Static Coverage	Dynamic Coverage
100	1,800	93.4%	94.5%
400	12,800	94.8%	97.5%
3,100	81,600	94.8%	98.1%
6,400	211,000	94.4%	98.9%

Table 1.4: Dynamic coverage of unseen text as a function of vocabulary size and amount of new text.

segments text into words) is needed. Then a unigram count for all of the spellings that occur in a corpus is determined. Those words that also occur in the dictionary are included. In addition a human screens the most frequent subset of new spellings to determine if they are words.

Table 1.3 shows the coverage of new text using a fixed vocabulary of a given size for English. For more inflectional languages such as French or German larger vocabulary sizes are required to achieve coverage similar to that of English. For a user of a speech recognition system, a more personalized vocabulary can be much more effective that a general fixed vocabulary. Table 1.4 shows the coverage as new words are added to a starting vocabulary of 20,000 words as more text is observed. In addition, Table 1.4 indicates the size of text recognized to add that many words. For many users, the dynamic coverage will be much better than what is shown in Table 1.4 with coverage ranging from 98.4% to 99.6% after 800 words are added.

1.6.4 Improved Language Models

A number of improvements have been proposed for the trigram LM. We give a brief overview of these models.

Class Models: Instead of using the actual words, one can use a set of word classes (which may be overlapping, i.e., a word may belong to many classes). Classes based on the part of speech tags, or the morphological analysis of words, or the semantic information have been tried. Also, automatically derived classes based on some statistical models of co-occurrence have been tried (see Brown, Della Pietra, et al., 1990). The general class model is:

$$p(W) = \sum_{c_1^n} \prod_{i=1}^{n} p(w_i|c_i)p(c_i|c_{i-2}, c_{i-1})$$

If the classes are non-overlapping, then c(w) is unique and the probability is:

$$p(W) = \prod_{i=1}^{n} p(w_i|c_i)p(c_i|c_{i-2}, c_{i-1})$$

These tri-class models have had higher perplexities that the corresponding trigram model. However, they have led to a reduction in perplexity when linearly combined with the trigram model.

Dynamic Models: Another idea introduced in DeMori and Kuhn (1990) is to take into account the document-long history to capture the burstiness of words. For example, in this section the probability that the word *model* will occur is much higher than its average frequency in general text. Using a cache of the recently observed words one can build a more dynamic LM using either the class model (DeMori & Kuhn, 1990) or the trigram model (Jelinek, Merialdo, et al., 1991). Expanding on this idea, one can can also affect the probability of related words called triggered words (see Lau, Rosenfeld, et al., 1993).

Mixture Models: Another approach is based on clustering corpora into several clusters. The linear combination of cluster-specific trigram models is used for modeling new text:

$$p(W) = \prod_{i=1}^{n} \sum_{j=1}^{k} \lambda_j p_j(w_n|w_{n-2}, w_{n-1})$$

where $p_j()$ is estimated from the j-th cluster of text. Another type of mixture is to use a sentence level mixture as in Iyer, Ostendorf, et al. (1994).

Structure-based Models: instead of using the most recent words' identity to define the equivalence class of a history, the state of a parser has been used to define the conditioning event (Goddeau & Zue, 1992). Also, the use of link grammar to capture long distance bigrams has been proposed recently (Lafferty, Sleator, et al., 1992).

1.6.5 Future Directions

There are several areas of research that can be pursued for improved language modeling.

- **Vocabulary Selection:** How to determine a vocabulary for a new domain particularly to personalize the vocabulary to a user while maximizing the coverage for a user's text. This is a problem that may be more severe for highly inflected languages and for the oriental languages where the notion of a word is not clearly defined for native speakers of the language.
- **Domain Adaptation:** How to estimate an effective language model for domains which may not have large online corpora of representative text. Another related problem is topic spotting where the topic-specific language model can be used to model the incoming text from a collection of domain-specific language models.
- **Incorporating Structure:** The current state-of-the-art in language modeling has not been able to improve on performance by the use of the structure (whether surface parse trees or the deep structure such as predicate argument structure) that is present in language. A concerted research effort to explore structure-based language model may be the key for a significant progress in language modeling. This will become more possible as annotated (parsed) data becomes available. Current research using probabilistic LR grammars, or probabilistic Context-Free grammars (including link grammars) is still in its infancy and would benefit from the increased availability of parsed data.

1.7 Speaker Recognition

Sadaoki Furui

NTT Human Interface Laboratories, Tokyo, Japan

1.7.1 Principles of Speaker Recognition

Speaker recognition, which can be classified into identification and verification, is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. This technique makes it possible to use the speaker's voice to verify their identity and control access to services such as voice dailing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers. AT&T and TI (with Sprint) have started field tests and actual application of speaker recognition technology; Sprint's Voice Phone Card is already being used by many customers. In this way, speaker recognition technology is expected to create new services that will make our daily lives more convenient. Another important application of speaker recognition technology is for forensic purposes.

Figure 1.8 shows the basic structures of speaker identification and verification systems. Speaker identification is the process of determining which registered speaker provides a given utterance. Speaker verification, on the other hand, is the process of accepting or rejecting the identity claim of a speaker. Most applications in which a voice is used as the key to confirm the identity of a speaker are classified as speaker verification.

There is also the case called *open set* identification, in which a reference model for an unknown speaker may not exist. This is usually the case in forensic applications. In this situation, an additional decision alternative, *the unknown does not match any of the models*, is required. In both verification and identification processes, an additional threshold test can be used to determine if the match is close enough to accept the decision or if more speech data are needed.

Speaker recognition methods can also be divided into text-dependent and text-independent methods. The former require the speaker to say key words or sentences having the same text for both training and recognition trials, whereas the latter do not rely on a specific text being spoken.

Both text-dependent and independent methods share a problem however. These systems can be easily deceived because someone who plays back the recorded voice of a registered speaker saying the key words or sentences can be accepted as the registered

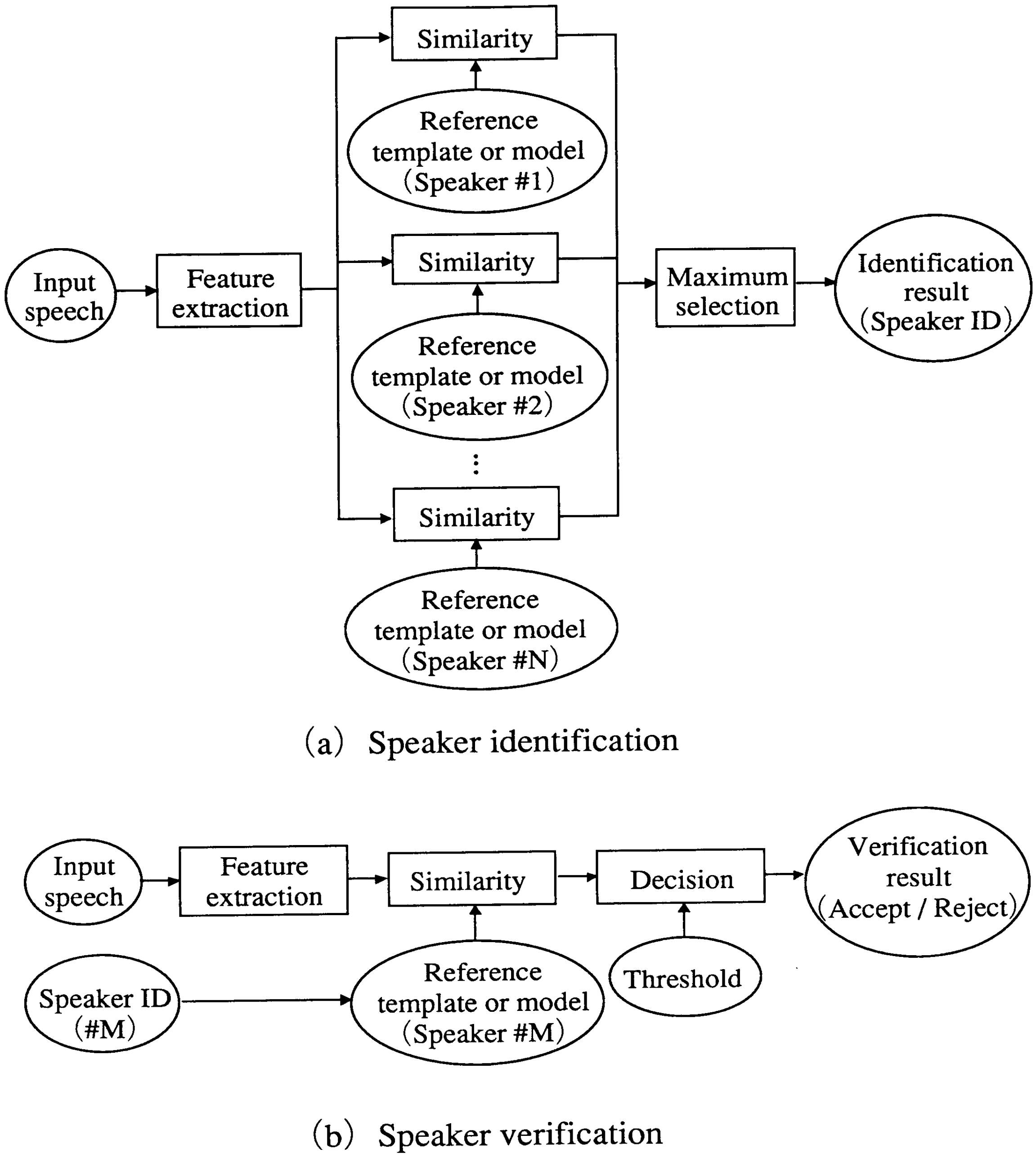

Figure 1.8: Basic structures of speaker recognition systems.

speaker. To cope with this problem, there are methods in which a small set of words, such as digits, are used as key words and each user is prompted to utter a given sequence of key words that is randomly chosen every time the system is used. Yet even this method is not completely reliable, since it can be deceived with advanced electronic recording equipment that can reproduce key words in a requested order. Therefore, a text-prompted (machine-driven-text-dependent) speaker recognition method has recently been proposed by Matsui and Furui (1993b).

1.7.2 Feature Parameters

Speaker identity is correlated with the physiological and behavioral characteristics of the speaker. These characteristics exist both in the spectral envelope (vocal tract characteristics) and in the supra-segmental features (voice source characteristics and dynamic features spanning several segments).

The most common short-term spectral measurements currently used are Linear Predictive Coding (LPC)-derived cepstral coefficients and their regression coefficients. A spectral envelope reconstructed from a truncated set of cepstral coefficients is much smoother than one reconstructed from LPC coefficients. Therefore it provides a stabler representation from one repetition to another of a particular speaker's utterances. As for the regression coefficients, typically the first- and second-order coefficients are extracted at every frame period to represent the spectral dynamics. These coefficients are derivatives of the time functions of the cepstral coefficients and are respectively called the delta- and delta-delta-cepstral coefficients.

1.7.3 Normalization Techniques

The most significant factor affecting automatic speaker recognition performance is variation in the signal characteristics from trial to trial (intersession variability and variability over time). Variations arise from the speaker themselves, from differences in recording and transmission conditions, and from background noise. Speakers cannot repeat an utterance precisely the same way from trial to trial. It is well known that samples of the same utterance recorded in one session are much more highly correlated than samples recorded in separate sessions. There are also long-term changes in voices.

It is important for speaker recognition systems to accommodate to these variations. Two types of normalization techniques have been tried; one in the parameter domain, and the other in the distance/similarity domain.

Parameter-Domain Normalization

Spectral equalization, the so-called *blind equalization* method, is a typical normalization technique in the parameter domain that has been confirmed to be effective in reducing linear channel effects and long-term spectral variation (Atal, 1974; Furui, 1981). This method is especially effective for text-dependent speaker recognition applications that use sufficiently long utterances. Cepstral coefficients are averaged over the duration of an entire utterance and the averaged values subtracted from the cepstral coefficients of each frame. Additive variation in the log spectral domain can be compensated for fairly well by this method. However, it unavoidably removes some text-dependent and speaker specific features; therefore it is inappropriate for short utterances in speaker recognition applications.

Distance/Similarity-Domain Normalization

A normalization method for distance (similarity, likelihood) values using a likelihood ratio has been proposed by Higgins, Bahler, et al. (1991). The likelihood ratio is defined as the ratio of two conditional probabilities of the observed measurements of the utterance: the first probability is the likelihood of the acoustic data given the claimed identity of the speaker, and the second is the likelihood given that the speaker is an imposter. The likelihood ratio normalization approximates optimal scoring in the Bayes sense.

A normalization method based on *a posteriori* probability has also been proposed by Matsui and Furui (1994a). The difference between the normalization method based on the likelihood ratio and the method based on *a posteriori* probability is whether or not the claimed speaker is included in the speaker set for normalization; the speaker set used in the method based on the likelihood ratio does not include the claimed speaker, whereas the normalization term for the method based on *a posteriori* probability is calculated by using all the reference speakers, including the claimed speaker.

Experimental results indicate that the two normalization methods are almost equally effective (Matsui & Furui, 1994a). They both improve speaker separability and reduce the need for speaker-dependent or text-dependent thresholding, as compared with scoring using only a model of the claimed speaker.

A new method in which the normalization term is approximated by the likelihood of a single mixture model representing the parameter distribution for all the reference speakers has recently been proposed. An advantage of this method is that the computational cost of calculating the normalization term is very small, and this method has been confirmed to give much better results than either of the above-mentioned normalization methods (Matsui & Furui, 1994a).

1.7.4 Text-Dependent Speaker Recognition Methods

Text-dependent methods are usually based on template-matching techniques. In this approach, the input utterance is represented by a sequence of feature vectors, generally short-term spectral feature vectors. The time axes of the input utterance and each reference template or reference model of the registered speakers are aligned using a dynamic time warping (DTW) algorithm and the degree of similarity between them, accumulated from the beginning to the end of the utterance, is calculated.

The hidden Markov model (HMM) can efficiently model statistical variation in spectral features. Therefore, HMM-based methods were introduced as extensions of the DTW-based methods, and have achieved significantly better recognition accuracies (Naik, Netsch, et al., 1989).

1.7.5 Text-Independent Speaker Recognition Methods

One of the most successful text-independent recognition methods is based on vector quantization (VQ). In this method, VQ codebooks consisting of a small number of representative feature vectors are used as an efficient means of characterizing speaker-specific features. A speaker-specific codebook is generated by clustering the training feature vectors of each speaker. In the recognition stage, an input utterance is vector-quantized using the codebook of each reference speaker and the VQ distortion accumulated over the entire input utterance is used to make the recognition decision.

Temporal variation in speech signal parameters over the long term can be represented by stochastic Markovian transitions between states. Therefore, methods using an ergodic HMM, where all possible transitions between states are allowed, have been proposed. Speech segments are classified into one of the broad phonetic categories corresponding to the HMM states. After the classification, appropriate features are selected.

In the training phase, reference templates are generated and verification thresholds are computed for each phonetic category. In the verification phase, after the phonetic categorization, a comparison with the reference template for each particular category provides a verification score for that category. The final verification score is a weighted linear combination of the scores from each category.

This method was extended to the richer class of mixture autoregressive (AR) HMMs. In these models, the states are described as a linear combination (mixture) of AR sources.

It can be shown that mixture models are equivalent to a larger HMM with simple states, with additional constraints on the possible transitions between states.

It has been shown that a continuous ergodic HMM method is far superior to a discrete ergodic HMM method and that a continuous ergodic HMM method is as robust as a VQ-based method when enough training data is available. However, when little data is available, the VQ-based method is more robust than a continuous HMM method (Matsui & Furui, 1993a).

A method using statistical dynamic features has recently been proposed. In this method, a multivariate auto-regression (MAR) model is applied to the time series of cepstral vectors and used to characterize speakers. It was reported that identification and verification rates were almost the same as obtained by an HMM-based method (Griffin, Matsui, et al., 1994).

1.7.6 Text-Prompted Speaker Recognition Method

In the text-prompted speaker recognition method, the recognition system prompts each user with a new key sentence every time the system is used and accepts the input utterance only when it decides that it was the registered speaker who repeated the prompted sentence. The sentence can be displayed as characters or spoken by a synthesized voice. Because the vocabulary is unlimited, prospective impostors cannot know in advance what sentence will be requested. Not only can this method accurately recognize speakers, but it can also reject utterances whose text differs from the prompted text, even if it is spoken by the registered speaker. A recorded voice can thus be correctly rejected.

This method is facilitated by using speaker-specific phoneme models as basic acoustic units. One of the major issues in applying this method is how to properly create these speaker-specific phoneme models from training utterances of a limited size. The phoneme models are represented by Gaussian-mixture continuous HMMs or tied-mixture HMMs, and they are made by adapting speaker-independent phoneme models to each speaker's voice. In order to properly adapt the models of phonemes that are not included in the training utterances, a new adaptation method based on tied-mixture HMMs was recently proposed by Matsui and Furui (1994b).

In the recognition stage, the system concatenates the phoneme models of each registered speaker to create a sentence HMM, according to the prompted text. Then the likelihood of the input speech matching the sentence model is calculated and used for the speaker recognition decision. If the likelihood is high enough, the speaker is accepted as the claimed speaker.

1.7.7 Future Directions

Although many recent advances and successes in speaker recognition have been achieved, there are still many problems for which good solutions remain to be found. Most of these problems arise from variability, including speaker-generated variability and variability in channel and recording conditions. It is very important to investigate feature parameters that are stable over time, insensitive to the variation of speaking manner, including the speaking rate and level, and robust against variations in voice quality due to causes such as voice disguise or colds. It is also important to develop a method to cope with the problem of distortion due to telephone sets and channels, and background and channel noises.

From the human-interface point of view, it is important to consider how the users should be prompted, and how recognition errors should be handled. Studies on ways to automatically extract the speech periods of each person separately from a dialogue involving more than two people have recently appeared as an extension of speaker recognition technology.

This section was not intended to be a comprehensive review of speaker recognition technology. Rather, it was intended to give an overview of recent advances and the problems which must be solved in the future. The reader is referred to the following papers for more general reviews: Furui, 1986a; Furui, 1989; Furui, 1991; Furui, 1994; O'Shaughnessy, 1986; Rosenberg & Soong, 1991.

1.8 Spoken Language Understanding[6]

Patti Price

SRI International, Menlo Park, California, USA

1.8.1 Overview

Spoken language understanding involves two primary component technologies (each covered elsewhere in this volume): speech recognition (SR), and natural language (NL) understanding. The integration of speech and natural language has great advantages: To NL, SR can bring prosodic information (information important for syntax and semantics but not well represented in text); NL can bring to SR additional knowledge sources (e.g., syntax and semantics). For both, integration affords the possibility of many more applications than could otherwise be envisioned, and the acquisition of new techniques and knowledge bases not previously represented. The integration of these technologies presents technical challenges, and challenges related to the quite different cultures, techniques and beliefs of the people representing the component technologies.

In large part, NL research has grown from symbolic systems approaches in computer science and linguistics departments. The desire to model language understanding is often motivated by a desire to understand cognitive processes, and therefore the underlying theories tend to be from linguistics and psychology. Practical applications have been less important than increasing intuitions about human processes. Therefore, coverage of phenomena of theoretical interest (usually the more rare phenomena) has traditionally been more important than broad coverage.

Speech recognition research, on the other hand, has largely been practiced in engineering departments. The desire to model speech is often motivated by a desire to produce practical applications. Techniques motivated by knowledge of human processes have therefore been less important than techniques that can be automatically developed or tuned, and broad coverage of a representative sample is more important than coverage of any particular phenomenon.

There are certainly technical challenges to the integration of SR and NL. However, progress toward meeting these challenges has been slowed by the differences outlined above. Collaboration can be inhibited by differences in motivation, interests, theoretical underpinnings, techniques, tools, and criteria for success. However, both groups have much to gain from collaboration. For the SR engineers, human language understanding

[6] I am grateful to Victor Zue for many very helpful suggestions.

provides an existence proof, and needs to be taken into account, since most applications involve interaction with at least one human. For the AI NL researchers, statistical and other engineering techniques can be important tools for their inquiries.

A survey of the papers on SR and NL in the last 5 to 10 years indicates that there is growing interest in the use of engineering techniques in NL investigations. Although the use of linguistic knowledge and techniques in engineering seems to have lagged, there are signs of growth as engineers tackle the more abstract linguistic units. These units are more rare, and therefore more difficult to model by standard, data-hungry engineering techniques.

1.8.2 State of the Art

Evaluation of spoken language understanding systems (see chapter 13) is required to estimate the state of the art objectively. However, evaluation itself has been one of the challenges of spoken language understanding. A brief survey of spoken language understanding work in the Europe, Japan and the U.S. is surveyed briefly below, and evaluation will be discussed in the following section.

Several sites in Canada, Europe and Japan have been researching spoken language understanding systems, including INRS in Canada, LIMSI in France, KTH in Sweden, the Center for Language Technology in Denmark, SRI International and DRA in the UK, Toshiba in Japan. The five year ESPRIT SUNDIAL project, which concluded in August 1993, involved several sites and the development of prototypes for train timetable queries in German and Italian and flight queries in English and French. All these systems are described in articles in Eurospeech (1993). The special issue of Speech Communication on Spoken Dialogue (Shirai & Furui, 1994), also includes several system descriptions, including those from NTT, MIT, Toshiba, and Canon.

In the ARPA program, the air travel planning domain has been chosen to support evaluation of spoken language systems (Pallett, 1991; Pallett, 1992; Pallett, Dahlgren, et al., 1992; Pallett, Fisher, et al., 1990; Pallett, Fiscus, et al., 1993; Pallett, Fiscus, et al., 1994; Pallett, Fiscus, et al., 1995). Vocabularies for these systems are usually about 2000 words. The speech and language are spontaneous, though fairly planned (since people are typically talking to a machine rather than to a person, and often use a push to talk button). The speech recognition utterance error rates in the December 1994 benchmarks was about 13% to 25%. The utterance understanding error rates range from 6% to 41%, although about 25% of the utterances are considered *unevaluable* in the testing paradigm, so these figures do not consider the same set (Pallett, 1991; Pallett, 1992; Pallett, Dahlgren, et al., 1992; Pallett, Fisher, et al., 1990; Pallett, Fiscus, et al., 1993; Pallett, Fiscus, et al., 1994; Pallett, Fiscus, et al., 1995). It may be that for limited domains, these error rates are compatible with many potential applications. Since conversational repairs in human-human dialogue can often be in the ranges observed for these systems, the bounding factor in applications may be not the error rates so much as the ability of the system to manage and recover from errors.

1.8.3 Evaluation of Spoken Language Understanding Systems

The benchmarks for spoken language understanding involve spontaneous speech input usually involving a real system, and sometimes with a human in the loop. The systems are scored in terms of the correctness of the response from the common database of information including flight and fare information. Performing this evaluation automatically requires human annotation to select the correct answer, define the minimal and maximal answers accepted, and to decide whether the query is ambiguous and/or answerable. The following sites participated in the most recent benchmarks for spoken language understanding: AT&T Bell Laboratories, Bolt Beranek and Newman, Carnegie Mellon University, Massachusetts Institute of Technology, MITRE, SRI International, and Unisys. Descriptions of these systems appear in ARPA (1995b).

There is a need to reduce the costs of evaluation, and to improve the quality of evaluations. One limitation of the current methodology is that the evaluated systems must be rather passive since the procedure does not generally allow for responses that are not a database response. This means that the benchmarks do not assess an important component of any real system: its ability to guide the user and to provide useful information in the face of limitations of the user or of the system itself. This aspect of the evaluation also forces the elimination of a significant portion of the data (about 25% in the most recent benchmark). Details on evaluation mechanisms are included in chapter 13. Despite the imperfections of these benchmarks, the sharing of ideas and the motivational aspects of the common benchmarks have yielded a great deal of technology transfer and communication.

1.8.4 Challenges

The integration of SR and NL in applications is faced with many of the same challenges that each of the components face: accuracy, robustness, portability, speed, and size, for example. However, the integration also gives rise to some new challenges as well, including: integration strategies, coordination of understanding components with system outputs, the effective use in NL of a new source of information from SR (prosody, in particular), and the handling of spontaneous speech effects (since people do not speak the way they write). Each of these areas will be described briefly below.

Integration

Several mechanisms for the communication among components have been explored. There is much evidence that human speech understanding involves the integration of a great variety of knowledge sources, including knowledge of the world or context, knowledge of the speaker and/or topic, lexical frequency, previous uses of a word or a semantically related topic, facial expressions, prosody, in addition to the acoustic attributes of the words. In SR, tighter integration of components has consistently led to improved performance, and tight integration of SR and NL has been a rather consistent goal. However, as grammatical coverage increases, standard NL techniques can become computationally difficult. Further, with increased coverage, NL tends to provide less constraint for SR.

The simplest approach of integration is simply to concatenate an existing speech recognition system and an existing NL system. However, this is suboptimal for several reasons. First, it is a very fragile interface and any errors that might be in the speech recognition system are propagated to the NL system. Second, the speech system does not then have a chance to take advantage of the more detailed syntactic, semantic and other higher level knowledge sources in deciding on what the words are. It is well known that people rely heavily on these sources in deciding what someone has said.

Perhaps the most important reason for the suboptimality of a simple concatenation is the fact that the writing mode differs greatly from the speaking mode. In the written form, people can create more complex sentences than in the spoken form because they have more time to think and plan. Readers have more time than do listeners to think and review, and they have visual cues to help ascertain the structure. Further, most instances of written text are not created in an interactive mode. Therefore, written communications tend to be more verbose than verbal communications. In non-interactive communications, the writer (or speaker in a non-interactive monologue) tries to foresee what questions a reader (or listener) may have. In an interactive dialogue, a speaker can usually rely on the other participant to ask questions when clarification is necessary, and therefore it is possible to be less verbose.

Another important difference between the written and spoken mode is that the spoken mode is strictly linear. A writer can pause for days or months before continuing a thought, can correct typos, can rearrange grammatical constructions and revise the organization of the material presented without leaving a trace in the result the reader sees. In spoken language interactions, every pause, restart, revision and hesitation has a consequence available to the listener. These effects are outlined further in the section below on spontaneous speech.

The differences between speaking and writing are compounded by the fact that most NL work has focussed on the written form, and if spoken language has been considered, except for rare examples such as Hindle (1983), it has largely been based on intuitions about the spoken language that would have occurred if not for the *noise* of spontaneous speech effects. As indicated in the overview, coverage of *interesting* linguistic phenomena has been a more important goal than testing coverage on occurring samples, written or spoken. More attention has been paid to correct analyses of complete sentences than to methods for recovery of interpretations when parses are incomplete (with the exceptions of some *robust* parsing techniques which still require a great deal more effort before they can be relied on in spoken language understanding systems (see section 3.7).

Because of the differences between speaking and writing, statistical models based on written materials will not match spoken language very well. Because of the fact that NL analyses have been predominantly based on complete parsing of *grammatically correct* sentences (based on intuitions of grammaticality of written text), traditional NL analyses often do very poorly when faced with transcribed spontaneous speech. Further, very little work has considered spontaneous effects. In sum, in general, simple concatenation of existing modules does not tend to work very well.

To combat the mismatch between existing SR and NL modules, two trends have been observed. The first is an increased use of *semantic* (as opposed to *syntactic grammars*) (see section 3.6). Such grammars rely on finding an interpretation without requiring *grammatical* input (where *grammatical* may be interpreted either in terms of traditional text-book grammaticality, or in terms of a particular grammar constructed for the task). Because semantic grammars focus on meaning in terms of the particular application, they can be more robust to grammatical deviations (see section 3.6). The second observed trend is the *n-best* interface. In the face of cultural and technical difficulties related to a tight integration, *n-best* integration has become popular. In this approach, the connection between SR and NL can be strictly serial: one component performs its computation, sends it to another component and that result is sent to yet another module. The inherent fragility of the strictly serial approach is mitigated by the fact that SR sends NL not just the best hypothesis from speech recognition, but the *n-best* (where N may be on the order of 10 to 100 sentence hypotheses). The NL component can then score hypotheses for grammaticality and/or use other knowledge sources to determine the best-scoring hypothesis. Frequently, the more costly knowledge sources are saved for this rescoring. More generally, there are several passes, a *progressive search* in which the search space is gradually narrowed and more knowledge sources are brought to bear. This approach is computationally tractable, and accommodates great modularity of design. The (D)ARPA, ESCA Eurospeech and ICSLP proceedings over the past several years contain several examples of the *n-best* approach and ways of bringing higher level knowledge sources to bear in SR (DARPA, 1990; DARPA, 1991a; DARPA, 1992a; ARPA, 1993a; ARPA, 1994; ARPA, 1995a; Eurospeech, 1989; Eurospeech, 1991; Eurospeech, 1993; ICSLP, 1990; ICSLP, 1992; ICSLP, 1994) . In addition, the special issue of Speech Communication on Spoken Dialogue (Shirai & Furui, 1994) contains several contributions investigating the integration of SR and NL.

Coordination of Understanding Components with System Outputs

With few exceptions, current research in spoken language systems has focused on the input side; i.e., the understanding of spoken input. However, many if not most potential applications involve a collaboration between the human and the computer. In many cases, spoken language output is an appropriate means of communication that may or may not be taken advantage of. Telephone-based applications are particularly important, since their use in spoken language understanding systems can make access to crucial data as convenient as the nearest phone, and since voice is the natural and (except for the as yet rare video-phones) usually the only modality available. Spoken outputs are also crucial in speech translation. The use of spoken output technologies, covered in more detail in chapter 5, is an important challenge to spoken language systems. In particular, we need reliable techniques to:

- decide when it is appropriate to provide a spoken output in conjunction with some other (e.g., screen-based) output and/or to instigate a clarification dialogue in order to recover from a potential misunderstanding,
- generate the content of spoken output given the data representation, context and dialogue state, and coordinate it with other outputs when present,
- synthesize a natural, easily interpreted and appropriate spoken version of the response taking advantage of the context and dialogue state to emphasize certain information or to express urgency, for example, and
- coordinate spoken outputs to guide the user toward usage better adapted to system capabilities.

Since people tend to be very cooperative in conversation, a system should not output structures it is not capable of understanding. By coordinating inputs and outputs the system can guide the user toward usage better adapted to the particular system. Not doing so can be very frustrating for the user.

Prosody

Prosody can be defined as the suprasegmental information in speech; that is, information that cannot be localized to a specific sound segment, or information that does not change the segmental identity of speech segments. For example, patterns of variation in fundamental frequency, duration, amplitude or intensity, pauses, and speaking rate have been shown to carry information about such prosodic elements as lexical stress, phrase breaks, and declarative or interrogative sentence form. Prosody consists of a phonological aspect (characterized by discrete, abstract units) and a phonetic aspect (characterized by continuously varying acoustic correlates).

Prosodic information is a source of information not available in text-based systems, except insofar as punctuation may indicate some prosodic information. Prosody can provide information about syntactic structure, it can convey discourse information, and it can also relay information about emotion and attitude. Surveys of how this can be done appear in Price and Ostendorf (1995); Shirai and Furui (1994); ESCA (1993).

Functionally, in languages of the world, prosody is used to indicate segmentation and saliency. The segmentation (or grouping) function of prosody may be related more to syntax (with some relation to semantics), while the saliency or prominence function may play a larger role in semantics than in syntax. To make maximum use of the potential of prosody will require tight integration, since the acoustic evidence needs to inform abstract units in syntax, semantics, discourse, and pragmatics.

Spontaneous Speech

The same acoustic attributes that indicate much of the prosodic structure (pitch and duration patterns) are also very common in aspects of spontaneous speech that seem to be more related to the speech planning process than to the structure of the utterance. For example, an extra long syllable followed by a pause can indicate either a large boundary that may be correlated with a syntactic boundary, or that the speaker is trying to plan the next part of the utterance. Similarly, a prominent syllable may mean that the syllable is new or important information, or that it replaces something previously said in error.

Disfluencies (e.g., *um*, repeated words, and repairs or false starts) are common in normal speech. It is possible that these phenomena can be isolated, e.g., by means of a posited *edit signal*, by joint modeling of intonation and duration, and/or by models that take into account syntactic patterns. However, modeling of speech disfluencies is only beginning to be modeled in spoken language systems. Two recent Ph.D. theses survey this topic (Lickley, 1994; Shriberg, 1994).

Disfluencies in human-human conversation are quite frequent, and a normal part of human communication. Their distribution is not random, and in fact may be a part of the communication itself. Disfluencies tend to be less frequent in human-computer interactions than in human-human interactions. However, the reduction in occurrences of disfluencies may be due to the fact that people are as yet not comfortable talking to computers. They may also be less frequent because there is more of an opportunity for the speaker to plan, and less of a potential for interruption. As people become increasingly comfortable with human-computer interactions and concentrate more on the task at hand than on monitoring their speech, disfluencies can be expected to increase. Speech disfluencies are a challenge to the integration of SR and NL since the evidence for disfluencies is distributed throughout all linguistic levels, from phonetic to at least the syntactic and semantic levels.

1.8.5 Future Directions

Although there have been significant recent gains in spoken language understanding, current technology is far from human-like: only systems in limited domains can be envisioned in the near term, and the portability of existing techniques is still rather limited. Application areas that appear to be a good match to technology on the near horizon include those that are naturally limited, for example database access (probably the most popular task across languages). With the rise in cellular phone use, and as rapid access to information becomes an increasingly important economic factor, telephone access to data and telephone transactions will no doubt rise dramatically. Mergers of telecommunications companies with video and computing companies will also no doubt add to the potential for automatic speech understanding.

While such short-term applications possibilities are exciting, if we can successfully meet the challenges outlined in previous sections, we can envision an information revolution on par with the development of writing systems. Spoken language is still the means of communication used first and foremost by humans, and only a small percentage of human communication is written. Automatic spoken language understanding can add to the many benefits of the spoken language many of the advantages normally associated only with text: random access, sorting, and access at different times and places. Making this vision a reality will require significant advances in the integration of SR and NL, and, in particular, the ability to better model prosody and disfluencies.

1.9 Chapter References

Acero, A. and Stern, R. M. (1990). Environmental robustness in automatic speech recognition. In *Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing*, pages 849–852, Albuquerque, New Mexico. Institute of Electrical and Electronic Engineers.

Alleva, F., Huang, X., and Hwang, M. Y. (1993). An improved search algorithm using incremental knowledge for continuous speech recognition. In *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 307–310, Minneapolis, Minnesota. Institute of Electrical and Electronic Engineers.

Alvarado, V. M. and Silverman, H. F. (1990). Experimental results showing the effects of optimal spacing between elements of a linear microphone array. In *Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing*, pages 837–840, Albuquerque, New Mexico. Institute of Electrical and Electronic Engineers.

Anastasakos, T., Makhoul, J., and Schwartz, R. (1994). Adaptation to new microphones using tied-mixture normalization. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 433–436, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Applebaum, T. H. and Hanson, B. A. (1989). Regression features for recognition of speech in quiet and in noise. In *Proceedings of the 1989 International Conference on Acoustics, Speech, and Signal Processing*, pages 985–988, Glasgow, Scotland. Institute of Electrical and Electronic Engineers.

ARPA (1993). *Proceedings of the 1993 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

ARPA (1994). *Proceedings of the 1994 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

ARPA (1995a). *Proceedings of the 1995 ARPA Human Language Technology Workshop*. Advanced Research Projects Agency, Morgan Kaufmann.

ARPA (1995b). *Proceedings of the ARPA Spoken Language Systems Technology Workshop*. Advanced Research Projects Agency, Morgan Kaufmann.

Atal, B. S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America*, 55(6):1304–1312.

Aubert, X., Dugast, C., Ney, H., and Steinbiss, V. (1994). Large vocabulary continuous speech recognition of wall street journal data. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 129–132, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Bahl, L. R., Bellegarda, J. R., de Souza, P. V., Gopalakrishnan, P. S., Nahamoo, D., and Picheny, M. A. (1993). Multonic Markov word models for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 1(3):334–344.

Bahl, L. R., Brown, P. F., de Souza, P. V., Mercer, R. L., and Picheny, M. A. (1993). A method for the construction of acoustic Markov models for words. *IEEE Transactions on Speech and Audio Processing*, 1(4):443–452.

Bahl, L. R., de Souza, P. V., Gopalakrishnan, P. S., Nahamoo, D., and Picheny, M. A. (1991). Decision trees for phonological rules in continuous speech. In *Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 185–188, Toronto. Institute of Electrical and Electronic Engineers.

Bahl, L. R., Jelinek, F., and Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):179–190.

Bellegarda, J. R., de Souza, P. V., Nadas, A. J., Nahamoo, D., Picheny, M. A., and Bahl, L. (1992). Robust speaker adaptation using a piecewise linear acoustic mapping. In *Proceedings of the 1992 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 445–448, San Francisco. Institute of Electrical and Electronic Engineers.

Bengio, Y., DeMori, R., Flammia, G., and Kompe, R. (1992). Global optimization of a neural network—hidden Markov model hybrid. *IEEE Transactions on Neural Networks*, 3(2):252–259.

Berger, A., Della Pietra, S., and Della Pietra, V. (1994). Maximum entropy methods in machine translation. Technical report, IBM Research Report.

Bocchieri, E. L. (1993). Vector quantization for the efficient computation of continuous density likelihoods. In *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 692–694, Minneapolis, Minnesota. Institute of Electrical and Electronic Engineers.

Brown, P. F., Della Pietra, V. J., de Souza, P. V., Lai, J. C., and Mercer, R. L. (1990). Class-based n-gram models of natural language. In *Proceedings of the IBM Natural Language ITL*, Paris, France.

Che, C., Lin, J., Pearson, J., de Vries, B., and Flanagan, J. (1994). Microphones arrays and neural networks for robust speech recognition. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Cohen, J., Gish, H., and Flanagan, J. (1994). Switchboard—the second year. Technical Report /pub/caipworks2 at ftp.rutgers.edu, CAIP Summer Workshop in Speech Recognition: Frontiers in Speech Processing II.

Cohen, J. R. (1989). Application of an auditory model to speech recognition. *Journal of the Acoustical Society of America*, 85(6):2623–2629.

Cole, R. A., Hirschman, L., et al. (1992). Workshop on spoken language understanding. Technical Report CSE 92-014, Oregon Graduate Institute of Science & Technology, P.O.Box 91000, Portland, OR 97291-1000 USA.

DARPA (1990). *Proceedings of the Third DARPA Speech and Natural Language Workshop*, Hidden Valley, Pennsylvania. Defense Advanced Research Projects Agency, Morgan Kaufmann.

DARPA (1991). *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, Pacific Grove, California. Defense Advanced Research Projects Agency, Morgan Kaufmann.

DARPA (1992). *Proceedings of the Fifth DARPA Speech and Natural Language Workshop*. Defense Advanced Research Projects Agency, Morgan Kaufmann.

Darroch, J. N. and Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43:1470–1480.

Daubechies, I. (1990). The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-36(5):961–1005.

Davis, S. B. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-28:357–366.

DeMori, R. and Kuhn, R. (1990). A cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-12(6):570–583.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum-likelihood from incompete data via the EM algorithm. *Journal of the Royal Statistical Society, Ser. B.*, 39:1–38.

Digalakis, V. and Murveit, H. (1994). Genones: Optimizing the degree of mixture tying in a large vocabulary hidden Markov model based speech recognizer. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 537–540, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Duda, R. O., Lyon, R. F., and Slaney, M. (1990). Correlograms and the separation of sounds. In *Proceedings of the 24th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 7457–7461.

Ephraim, Y. (1992). Gain-adapted hidden Markov models for recognition of clean and noisy speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 40:1303–1316.

Erell, A. and Weintraub, M. (1990). Recognition of noisy speech: Using minimum-mean log-spectral distance estimation. In *Proceedings of the Third DARPA Speech and Natural Language Workshop*, pages 341–345, Hidden Valley, Pennsylvania. Defense Advanced Research Projects Agency, Morgan Kaufmann.

ESCA (1993). Proceedings of the ESCA workshop on prosody. Technical Report Working Papers 41, Lund University Department of Linguistics.

Eurospeech (1989). *Eurospeech '89, Proceedings of the First European Conference on Speech Communication and Technology*, Paris. European Speech Communication Association, European Speech Communication Association.

Eurospeech (1991). *Eurospeech '91, Proceedings of the Second European Conference on Speech Communication and Technology*, Genova, Italy. European Speech Communication Association.

Eurospeech (1993). *Eurospeech '93, Proceedings of the Third European Conference on Speech Communication and Technology*, Berlin. European Speech Communication Association.

Fanty, M., Barnard, E., and Cole, R. A. (1995). Alphabet recognition. In *Handbook of Neural Computation*. Publisher Unknown. In press.

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(2):254–272.

Furui, S. (1986a). Research on individuality features in speech waves and automatic speaker recognition techniques. *Speech Communication*, 5(2):183–197.

Furui, S. (1986b). Speaker-independent isolated word recognition using dynamic features of the speech spectrum. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(1):59–59.

Furui, S. (1989). *Digital Speech Processing, Synthesis, and Recognition*. Marcel Dekker, New York.

Furui, S. (1991). Speaker-dependent-feature extraction, recognition and processing techniques. *Speech Communication*, 10(5-6):505–520.

Furui, S. (1994). An overview of speaker recognition technology. In *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pages 1–9.

Gales, M. J. F. and Young, S. J. (1992). An improved approach to the hidden Markov model decomposition of speech and noise. In *Proceedings of the 1992 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 233–236, San Francisco. Institute of Electrical and Electronic Engineers.

Gauvain, J.-L. and Lee, C.-H. (1991). Bayesian learning for hidden markov model with gaussian mixture state observation densities. In *Eurospeech '91, Proceedings of the Second European Conference on Speech Communication and Technology*, pages 939–942, Genova, Italy. European Speech Communication Association.

Ghitza, O. (1988). Temporal non-place information in the auditory-nerve firing patterns as a front end for speech recognition in a noisy environment. *Journal of Phonetics*, 16(1):109–124.

Goddeau, D. and Zue, V. (1992). Integrating probabilistic LR parsing into speech understanding systems. In *Proceedings of the 1992 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184, San Francisco. Institute of Electrical and Electronic Engineers.

Greenberg, S. (1988). Theme issue: Representation of speech in the auditory periphery. *Journal of Phonetics*, 16(1).

Griffin, C., Matsui, T., and Furui, S. (1994). Distance measures for text-independent speaker recognition based on MAR model. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 309–312, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Haeb-Umbach, R., Geller, D., and Ney, H. (1993). Improvements in connected digit recognition using linear discriminant analysis and mixture densities. In *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 239–242, Minneapolis, Minnesota. Institute of Electrical and Electronic Engineers.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis for speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752.

Hermansky, H., Morgan, N., Bayya, A., and Kohn, P. (1991). Compensation for the effects of the communication channel in auditory-like analysis of speech. In *Eurospeech '91, Proceedings of the Second European Conference on Speech Communication and Technology*, pages 1367–1370, Genova, Italy. European Speech Communication Association.

Hermansky, H., Morgan, N., and Hirsch, H. G. (1993). Recognition of speech in additive and convolutional noise based on RASTA spectral processing. In *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 83–86, Minneapolis, Minnesota. Institute of Electrical and Electronic Engineers.

Higgins, A. L., Bahler, L., and Porter, J. (1991). Speaker verification using randomized phrase prompting. *Digital Signal Processing*, 1:89–106.

Hindle, D. (1983). Deterministic parsing of syntactic nonfluencies. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, pages 123–128, Cambridge, Massachusetts. Association for Computational Linguistics.

Hirsch, H. G., Meyer, P., and Ruehl, H. W. (1991). Improved speech recognition using high-pass filtering of subband envelopes. In *Eurospeech '91, Proceedings of the Second European Conference on Speech Communication and Technology*, pages 413–416, Genova, Italy. European Speech Communication Association.

Hon, H.-W. and Lee, K.-F. (1991). CMU robust vocabulary-independent speech recognition system. In *Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 889–892, Toronto. Institute of Electrical and Electronic Engineers.

Huang, X. D., Ariki, Y., and Jack, M. (1990). *Hidden Markov Models for Speech Recognition*. Edinburgh University Press.

Huang, X. D. and Lee, K. F. (1993). On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition. *IEEE Transactions on Speech and Audio Processing*, 1(2):150–157.

Hunt, M. J. (1993). Signal processing for speech. In Asher, R. E., editor, *The Encyclopedia of Language and Linguistics*. Pergamon Press.

Hunt, M. J. and Lefèbvre, C. (1989). A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. In *Proceedings of the 1989 International Conference on Acoustics, Speech, and Signal Processing*, pages 262–265, Glasgow, Scotland. Institute of Electrical and Electronic Engineers.

Hwang, M. Y. and Huang, X. (1993). Shared-distribution hidden Markov models for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 1(4):414–420.

ICASSP (1987). *Proceedings of the 1987 International Conference on Acoustics, Speech, and Signal Processing*, Dallas. Institute of Electrical and Electronic Engineers.

ICASSP (1989). *Proceedings of the 1989 International Conference on Acoustics, Speech, and Signal Processing*, Glasgow, Scotland. Institute of Electrical and Electronic Engineers.

ICASSP (1990). *Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing*, Albuquerque, New Mexico. Institute of Electrical and Electronic Engineers.

ICASSP (1991). *Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing*, Toronto. Institute of Electrical and Electronic Engineers.

ICASSP (1992). *Proceedings of the 1992 International Conference on Acoustics, Speech, and Signal Processing*, San Francisco. Institute of Electrical and Electronic Engineers.

ICASSP (1993). *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, Minneapolis, Minnesota. Institute of Electrical and Electronic Engineers.

ICASSP (1994). *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

ICSLP (1990). *Proceedings of the 1990 International Conference on Spoken Language Processing*, Kobe, Japan.

ICSLP (1992). *Proceedings of the 1992 International Conference on Spoken Language Processing*, Banff, Alberta, Canada. University of Alberta.

ICSLP (1994). *Proceedings of the 1994 International Conference on Spoken Language Processing*, Yokohama, Japan.

Iyer, R., Ostendorf, M., and Rohlicek, R. (1994). An improved language model using a mixture of Markov components. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Jelinek, F. (1969). A fast sequential decoding algorithm using a stack. *IBM journal of Research and Development*, 13.

Jelinek, F., Merialdo, B., Roukos, S., and Strauss, M. (1991). A dynamic language model for speech recognition. In *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, pages 293–295, Pacific Grove, California. Defense Advanced Research Projects Agency, Morgan Kaufmann.

Juang, B. H. (1991). Speech recognition in adverse environments. *Computer Speech and Language*, pages 275–294.

Juang, B. H., Rabiner, L. R., and Wilpon, J. G. (1986). On the use of bandpass liftering in speech recognition. In *Proceedings of the 1986 International Conference on Acoustics, Speech, and Signal Processing*, pages 765–768, Tokyo. Institute of Electrical and Electronic Engineers.

Koehler, J., Morgan, N., Hermansky, H., Hirsch, H. G., and Tong, G. (1994). Integrating RASTA-PLP into speech recognition. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 421–424, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Kubala, F., Anastasakos, A., Makhoul, J., Nguyen, L., Schwartz, R., and Zavaliagkos, G. (1994). Comparative experiments on large vocabulary speech recognition. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 561–564, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Kuhn, R., De Mori, R., and Millien, E. (1994). Learning consistent semantics from training data. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 37–40, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Lafferty, J., Sleator, D., and Temperley, D. (1992). Grammatical trigrams: A probabilistic model of link grammar. In *Proceedings of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language.*

Lau, R., Rosenfeld, R., and Roukos, S. (1993). Trigger-based language models: A maximum entropy approach. In *Proceedings of the 1993 International Conference* *on Acoustics, Speech, and Signal Processing*, volume 2, pages 45–48, Minneapolis, Minnesota. Institute of Electrical and Electronic Engineers.

Lickley, R. J. (1994). *Detecting Disfluency in Spontaneous Speech.* PhD thesis, University of Edinburgh, Scotland.

Lim, J. and Oppenheim, A. (1979). Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67:1586–1604.

Lippmann, R. P., Martin, F. A., and Paul, D. B. (1987). Multi-style training for robust isolated-word speech recognition. In *Proceedings of the 1987 International Conference on Acoustics, Speech, and Signal Processing*, pages 709–712, Dallas. Institute of Electrical and Electronic Engineers.

Liu, F.-H., Stern, R. M., Acero, A., and Moreno, P. (1994). Environment normalization for robust speech recognition using direct cepstral comparison. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 61–64, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Lockwood, P., Boudy, J., and Blanchet, M. (1992). Non-linear spectral subtraction (NSS) and hidden Markov models for robust speech recognition in car noise environments. In *Proceedings of the 1992 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 265–268, San Francisco. Institute of Electrical and Electronic Engineers.

Lyon, R. F. (1982). A computational model of filtering, detection, and compression in the cochlea. In *Proceedings of the 1982 International Conference on Acoustics, Speech, and Signal Processing*, pages 1282–1285. Institute of Electrical and Electronic Engineers.

Markel, J. D. and Gray, Jr., A. H. (1976). *Linear Prediction of Speech.* Springer-Verlag, Berlin.

Matsui, T. and Furui, S. (1993a). Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs. In *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 157–160, Minneapolis, Minnesota. Institute of Electrical and Electronic Engineers.

Matsui, T. and Furui, S. (1993b). Concatenated phoneme models for text-variable speaker recognition. In *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 391–394, Minneapolis, Minnesota. Institute of Electrical and Electronic Engineers.

Matsui, T. and Furui, S. (1994a). Similarity normalization method for speaker verification based on a posteriori probability. In *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pages 59–62.

Matsui, T. and Furui, S. (1994b). Speaker adaptation of tied-mixture-based phoneme models for text-prompted speaker recognition. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 125–128, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Meng, H. M. and Zue, V. W. (1990). A comparative study of acoustic representations of speech for vowel classification using multi-layer perceptrons. In *Proceedings of the 1990 International Conference on Spoken Language Processing*, volume 2, pages 1053–1056, Kobe, Japan.

Merhav, N. and Ephraim, Y. (1991). Maximum likelihood hidden markov modeling using a dominant state sequence of states. *IEEE Transactions on Signal Processing*, 39(9):2111–2114.

Murveit, H., Butzberger, J., Digilakis, V., and Weintraub, M. (1993). Large-vocabulary dictation using SRI's DECIPHER speech recognition system: Progressive search techniques. In *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 319–322, Minneapolis, Minnesota. Institute of Electrical and Electronic Engineers.

Naik, J. M., Netsch, L. P., and Doddington, G. R. (1989). Speaker verification over long distance telephone lines. In *Proceedings of the 1989 International Conference on Acoustics, Speech, and Signal Processing*, pages 524–527, Glasgow, Scotland. Institute of Electrical and Electronic Engineers.

Neumeyer, L. and Weintraub, M. (1994). Probabilistic optimum filtering for robust speech recognition. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 417–420, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Ney, H., Mergel, D., Noll, A., and Paesler, A. (1992). Data driven search organization for continuous speech recognition. *IEEE Transactions on Signal Processing*, 40(2):272–281.

Nilsson, N. J. (1971). *Problem-Solving Methods in Artificial Intelligence.* McGraw-Hill, New York.

Ohshima, Y. (1993). *Robustness in Speech Recognition using Physiologically-Motivated Signal Processing.* PhD thesis, CMU.

O'Shaughnessy, D. (1986). Speaker recognition. *IEEE Acoustics, Speech and Signal Processing Magazine*, 3(4):4–17.

Pallett, D. (1991). DARPA resource management and ATIS benchmark test poster session. In *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, pages 49–58, Pacific Grove, California. Defense Advanced Research Projects Agency, Morgan Kaufmann.

Pallett, D. (1992). ATIS benchmarks. In *Proceedings of the Fifth DARPA Speech and Natural Language Workshop*. Defense Advanced Research Projects Agency, Morgan Kaufmann.

Pallett, D., Dahlgren, N., Fiscus, J., Fisher, W., Garofolo, J., and Tjaden, B. (1992). DARPA February 1992 ATIS benchmark test results. In *Proceedings of the Fifth DARPA Speech and Natural Language Workshop*, pages 15–27. Defense Advanced Research Projects Agency, Morgan Kaufmann.

Pallett, D., Fiscus, J., Fisher, W., and Garofolo, J. (1993). Benchmark tests for the DARPA spoken language program. In *Proceedings of the 1993 ARPA Human Language Technology Workshop*, pages 7–18, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Pallett, D., Fiscus, J., Fisher, W., Garofolo, J., Lund, B., and Prysbocki, M. (1994). 1993 benchmark tests for the ARPA spoken language program. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*, pages 49–74, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Pallett, D., Fisher, W., Fiscus, J., and Garofolo, J. (1990). DARPA ATIS test results. In *Proceedings of the Third DARPA Speech and Natural Language Workshop*, pages 114–121, Hidden Valley, Pennsylvania. Defense Advanced Research Projects Agency, Morgan Kaufmann.

Pallett, D. S., Fiscus, J. G., Fisher, W. M., Garofolo, J. S., Lund, B. A., Martin, A., and Przybocki, M. A. (1995). 1994 benchmark tests for the ARPA spoken language program. In *Proceedings of the 1995 ARPA Human Language Technology Workshop*, pages 5–36. Advanced Research Projects Agency, Morgan Kaufmann.

Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand, M. (1991). Complex sounds and auditory images. In *Auditory Physiology and Perception*, pages 429–446. Pergamon Press.

Paul, D. B. (1994). The Lincoln large-vocabulary stack-decoder based HMM CSR. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*, pages

374–379, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Peterson, P. M. (1989). Adaptive array processing for multiple microphone hearing aids. Technical Report 541, Research Laboratory of Electronics, MIT, Cambridge, Massachusetts.

Porter, J. E. and Boll, S. F. (1984). Optimal estimators for spectral restoration of noisy speech. In *Proceedings of the 1984 International Conference on Acoustics, Speech, and Signal Processing*, pages 18.A.2.1–4. Institute of Electrical and Electronic Engineers.

Price, P. and Ostendorf, M. (1995). Combining linguistic with statistical methods in modeling prosody. In Morgan, J. L. and Demuth, K., editors, *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Rabiner, L. R. and Schafer, R. W. (1978). *Digital Processing of Speech Signals*. Signal Processing. Prentice-Hall, Englewood Cliffs, New Jersey.

Rosenberg, A. E. and Soong, F. K. (1991). Recent research in automatic speaker recognition. In Furui, S. and Sondhi, M. M., editors, *Advances in Speech Signal Processing*, pages 701–737. Marcel Dekker, New York.

Schroeter, J. and Sondhi, M. M. (1994). Techniques for estimating vocal tract shapes from the speech signal. *IEEE Transactions on Speech and Audio Processing*, 2(1):133–150.

Schwartz, R., Chow, Y., and Kubala, F. (1987). Rapid speaker adaption using a probabalistic spectral mapping. In *Proceedings of the 1987 International Conference on Acoustics, Speech, and Signal Processing*, pages 633–636, Dallas. Institute of Electrical and Electronic Engineers.

Seneff, S. (1988). A joint synchrony/mean-rate model of auditory speech processing. *Journal of Phonetics*, 16(1):55–76.

Shirai, K. and Furui, S. (1994). Special issue on spoken dialogue. *Speech Communication*, 15(3-4).

Shriberg, E. E. (1994). *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis, Stanford University.

Shukat-Talamazzini, E. G., Niemann, H., Eckert, W., Kuhn, T., and Rieck, S. (1992). Acoustic modeling of sub-word units in the ISADORA speech recognizer. In *Proceedings of the 1992 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 577–580, San Francisco. Institute of Electrical and Electronic Engineers.

Stockham, T. G., J., Connon, T. M., and Ingebretsen, R. B. (1975). Blind deconvolution through digital signal processing. *Proceedings of the IEEE*, 63(4):678–692.

Sullivan, T. M. and Stern, R. M. (1993). Multi-microphone correlation-based processing for robust speech recognition. In *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 91–94, Minneapolis, Minnesota. Institute of Electrical and Electronic Engineers.

Van Compernolle, D. (1990). Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings. In *Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing*, pages 833–836, Albuquerque, New Mexico. Institute of Electrical and Electronic Engineers.

Varga, A. P. and Moore, R. K. (1990). Hidden Markov model decomposition of speech and noise. In *Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing*, pages 845–848, Albuquerque, New Mexico. Institute of Electrical and Electronic Engineers.

Waibel, A. and Lee, K. F. (1990). *Readings in Speech Recognition*. Morgan Kaufmann.

Zue, V., Glass, J., Phillips, M., and Seneff, S. (1990). The MIT SUMMIT speech recognition system: A progress report. In *Proceedings of the Third DARPA Speech and Natural Language Workshop*, Hidden Valley, Pennsylvania. Defense Advanced Research Projects Agency, Morgan Kaufmann.