A Survey on Automatic Speech Recognition with an Illustrative Example on Continuous Speech Recognition of Mandarin

Chin-Hui Lee*, Biing-Hwang Juang*

Abstract

For the past two decades, research in speech recognition has been intensively carried out worldwide, spurred on by advances in signal processing, algorithms, architectures, and hardware. Speech recognition systems have been developed for a wide variety of applications, ranging from small vocabulary keyword recognition over dial-up telephone lines, to medium size vocabulary voice interactive command and control systems on personal computers, to large vocabulary speech dictation, spontaneous speech understanding, and limited-domain speech translation. In this paper we review some of the key advances in several areas of automatic speech recognition. We also illustrate, by examples, how these key advances can be used for continuous speech recognition of Mandarin. Finally we elaborate the requirements in designing successful real-world applications and address technical challenges that need to be harnessed in order to reach the ultimate goal of providing an easy-to-use, natural, and flexible voice interface between people and machines.

Keyword: hidden Markov modeling, dynamic programming, speech recognition, acoustic modeling, Mandarin speech recognition, spoken language systems

1. Introduction

In the past few years a significant portion of the research in speech processing has gone into studying practical methods for automatic speech recognition (ASR). Much of this effort has been stimulated by the Advanced Research Project Agency (ARPA), formerly known as D(efense)ARPA, which has funded research on three large vocabulary recognition (LVR) projects, namely the Naval Resource Management (RM) task, the Air Travel Information System (ATIS) and the North American Business (NAB, previously

---

* Multimedia Communications Research Lab, Lucent Technologies, 600 Mountain Ave., Murray Hill, Bell Laboratories, NJ 07974, U.S.A.
   E-mail: {chl, bhj}@research.bell-labs.com

known as the Wall Street Journal or WSJ) task. In addition, there is a worldwide activity in multi-lingual, large vocabulary speech recognition because of the potential applications to voice-interactive database access and management (e.g. ATIS & RM), voice dictation (e.g. discrete word recognizer [Jelinek 1985] and continuous speech recognition such as the NAB/WSJ task) and limited-domain spoken language translation. The Philips SPICOS system and its extensions, the CSELT system (which is currently in trial) for Eurorail information services, the Cambridge University systems, and the LIMSI effort, are examples of the current activity in speech recognition research in Europe. In Japan, large vocabulary recognition systems are being developed based on the concept of *interpreting telephony* and telephone directory assistance. In Taiwan and China, syllable-based recognizers have been designed to handle large vocabulary Mandarin dictation which is of practical importance because keyboard entry of Chinese text requires a considerable amount of effort and training (e.g. [Lee *et al.* 1993]. In Canada, the most notable research project is the INRS 86,000-word isolated word recognition system. In the United States, in addition to the research being carried out at AT&T and IBM, most of the effort is sponsored by ARPA, encompassing efforts by BBN (the BYBLOS system), CMU (the SPHINX systems), Dragon, Lincoln Laboratory, MIT (the Summit system and its extensions), SRI (the DECIPHER system), and many others in the ARPA Human Language Technology Program. A brief history of automatic speech recognition research can be found in the textbook on speech recognition by Rabiner and Juang (1993).

Although we have learned a great deal about how to build practical and useful speech recognition systems, there remain a number of fundamental questions about the technology to which we have no definitive answers. It is clear that the speech signal is one of the most complex signals that we need to deal with. It is produced by a human's vocal system and therefore not easy to be characterized by a simple 2-dimensional model of sound propagation. While there exist a number of sophisticated mathematical models which attempt to simulate the speech production system, their modeling capability is still limited. Some of these models can be found in the seminal text by Flanagan (1964). In addition to the inherent physiological complexity of the human vocal tract, the physical production system differs from one person to another. The speech signal being observed is different (even when produced by the same person) each time, even for multiple utterances of the same sequence of words. Part of the reason that automatic speech recognition by machine is difficult is due to this inherent signal variability. In addition to the vast inherent differences across different speakers and different dialects, the speech signal is influenced by the transducer used to capture the signal, the channel used to transmit the signal, and the speaking environment that can add noise to the speech signal

or change the way the signal is produced (e.g. the *Lombard effect* shown in [Junqua *et al*. 1993]) in very noisy environments.

There have been many attempts to find so called *distinctive features* of speech (e.g. [Fant 1973]) which are invariant to a number of factors. Certain distinctive (phonetic) features, such as nasality and voicing, can be used to represent the place and manner of articulation of speech sounds so that speech can be uniquely identified by detecting the acoustic-phonetic properties of the signal. By organizing such knowledge in a systematic manner, speech recognition can (in theory) be performed by first identifying and labeling the sequence of feature vectors and then identifying the corresponding sounds in the speech signal, followed by decoding the corresponding sequence of words using lexical access to a dictionary of words. This has been demonstrated in spectrogram reading by a human expert who can visually segment and identify some speech sounds based on knowledge of acoustic-phonetics of English. Although the collection of distinctive features, in theory, offers a set of *invariant* features for speech recognition, it is not generally used in most speech recognitions systems. This is due to the fact that the set of distinctive features are usually difficult to identify in spontaneous continuous speech and the recognition results are generally unreliable.

A more successful approach to automatic speech recognition is to treat the speech signal as a stochastic pattern and to adopt a statistical pattern recognition approach. For this approach we assume a source-channel speech generation model (e.g. [Bahl *et al*. 1983]) shown in Figure 1, in which the source produces a sequence of words, W. Because of uncertainty and inaccuracy in converting from words to speech, we model the conversion from W to an observed speech waveform, S, as a noisy channel. Speech recognition is then formulated as a *maximum a posteriori* (MAP) decoding problem, as shown in Figure 1. Instead of working with the speech signal S directly, one way to simplify the problem is to assume that S is first parametrically represented as a sequence of acoustic vectors A. We then use the Bayes rule to reformulate the decoding problem as follows,

$$(1) \qquad \underset{W \in \Gamma}{\mathrm{argmax}} \; P(W|A) = \underset{W \in \Gamma}{\mathrm{argmax}} \; P(A|W) \cdot P(W) \; ,$$

where Γ is the set of all possible sequences of words, $P(A|W)$ is the conditional probability of the acoustic vector sequence, A, given a particular sequence of words W, and $P(W)$ is the a priori probability of generating the sequence of words W. The first term, $P(A|W)$, is often referred to as an *acoustic model*, and the second term, $P(W)$, is known as a language model. The noisy channel in Figure 1 is a model jointly characterizing the speech

production system, the speaker variability, the speaking environment, and the trans-mission medium. Since it is not feasible to have a complete knowledge about such a noisy channel, the statistical approach often assumes particular parametric forms for P_θ $(A|W)$ and $P_\omega(W)$, i.e. according to specific models. All the parameters of the statistical models (i.e. θ and ω) needed in evaluting the acoustic probability, $P_\theta(W|A)$, and the language probability, $P_\omega(W)$, are usually estimated from a large collection (the so-called *training set*) of speech and text training data. This process is often referred to as *model training* or *learning*. We will discuss this important issue later in the paper.

There is some recent attempt trying to separate the speech production part from the source-channel model by incorporating knowledge about the human speech production mechanism. Knowledge about the transducers used for capturing speech and the channel used for transmitting speech can also be explicitly modeled. However, the effectiveness of such approaches is yet to be shown.
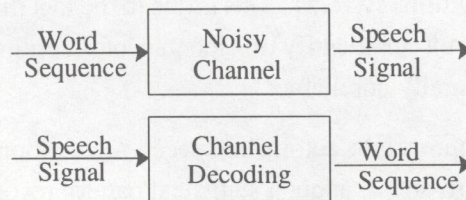
Figure 1 Source-channel model of speech generation and
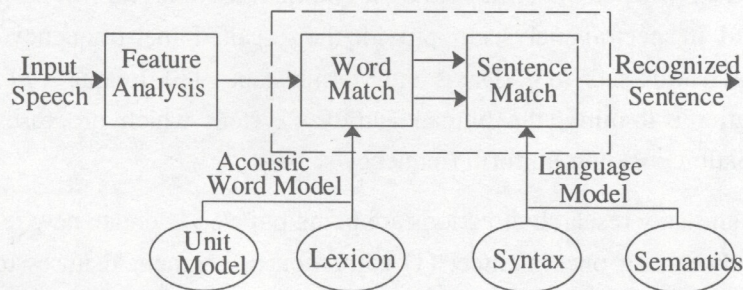speech recognition

In the following sections we first briefly review the statistical pattern recognition approach to speech recognition. We then describe the two most important techniques that have helped to advance the state of the art of automatic speech recognition, namely *hidden Markov modeling* (HMM) of the speech signal and *dynamic programming* (DP) methods for best path decoding of structural lexical networks. We next discuss several ASR systems and some real-world applications. Finally we address ASR system design considerations and present a number of ASR research challenges we need to overcome in order to deploy natural human-machine interactive speech input/output systems.

2. Pattern Recognition Approach

A block diagram of an integrated approach to continuous speech recognition is shown in Figure 2. The feature analysis module provides the acoustic feature vectors used to characterize the spectral properties of the time varying speech signal. The word-level

acoustic match module evaluates the similarity between the input feature vector sequence (corresponding to a portion of the input speech) and a set of acoustic word models for all words in the recognition task vocabulary to determine which words were most likely spoken. The sentence-level match module uses a language model (i.e. a model of syntax and semantics) to determine the most likely sequence of words. Syntactic and semantic rules can be specified either manually, based on task constraints, or with statistical models such as word and class *N*-gram probabilities. Search and recognition decisions are made by considering all likely word sequences and choosing the one with the best overall matching score as the recognized sentence.

Figure 2 *Block diagram of a typical integrated continuous speech recognizer*

2.1 Speech Analysis and Feature Extraction

The purpose of the feature analysis module is to parametrize the speech into a parsimonious sequence of feature vectors that contain the relevant (for recognition) information about the sounds within the utterance. Although there is no consensus as to what constitutes the optimal feature analysis, most systems extract spectral features with the following properties: having good discrimination to readily distinguish between similar speech sounds, being easy to model statistically without the need for an excessive amount of training data, and having statistical properties which are somewhat invariant across speakers and over a wide range of speaking environments. To our knowledge there is no single feature set that possesses all the above properties. The features used in speech recognition systems are largely derived from their utility in speech analysis, speech coding, and psycho-acoustics.

Fourier analysis is still the most widely used method for extracting spectral features for speech recognition. Implementations of feature extraction include:

- *Short-Time Spectral Features*: Most recognition systems use either *discrete*

Fourier transform (DFT) or *linear predictive coding* (LPC) spectral analysis methods based on fixed size frames of windowed speech data, and extract spectral features, including *LPC-derived features*, such as *reflection coefficients, log area ratios, line spectral frequencies, composite sinusoidal model parameters, auto-correlations,* and *cumulant features*. In the last few years the set of short-time spectral feature sets for each frame has been extended to include *dynamic* information (e.g. the first and the second order derivatives) of the features. The most popular such representation includes *cepstral* features along with its first and second time derivatives (e.g. [Furui 1986]).

- *Frequency-Warped Spectral Features*: Sometimes non-uniform frequency scales are used in spectral analysis to provide the so-called mel-frequency or bark-scale spectral feature sets (e.g. [Davis and Mermelstein 1980, Junqua *et al.* 1993]). The motivation is to mimic the human auditory system which processes the spectral information on a non-uniform frequency scale.

Although many research directions are being pursued to create new feature sets for ASR, most promising ones include: (1) extraction of segment features to allow variable-length speech analysis to incorporate discriminative information from neighboring frames; (2) extraction of articulatory and auditory features to make use of the human speech production and perception mechanism (e.g. [Deng and Sun 1994, Ghitza 1988, Seneff 1988]); (3) extraction of discriminative features which are functions of the data, the task, and the classifiers being used (e.g. [Ney *et al.* 1992, Biem *et al.* 1993, Rahim and Lee 1996]).

2.2 Selection of Fundamental Speech Units

The word-level acoustic match module determines the optimal word match based on a set of subword models and a lexicon. The subword models are the building blocks for words, phrases, and sentences. Ideally, subword models must be easy to train from a finite set of speech material and robust to natural variations in accent, word pronunciation, etc., and provide high recognition accuracy for the intended task.

Subword units corresponding to phonetic classes are used in most speech recognition systems today. Such units are modeled acoustically based on a lexical description of the words in the training set. In general, no assumption is made, *a priori*, about the mapping between acoustic measurements and subword linguistic units. This mapping is entirely learned via a finite training set of speech utterances. The resulting units, which we call *phoneme-like units* or PLUs, are essentially acoustic models of linguistically-based units as *represented in the words occurring in the given training*

set. Since the set of PLUs are usually chosen and designed to cover all the phonetic labels of a particular language, and words in the language can usually be pronounced based on this set of fundamental speech units, this pattern recognition approach offers the potential of modeling virtually all the words and word sequences in the language.

The simplest set of fundamental speech units are phones that correspond to the basic phonemes of the language.These basic speech units are often called context-independent PLUs since the sounds are represented independent of the linguistic context in which they occur. Other choices for subword units include:

- *Units Other than Phones*: Units smaller than a phone, such as phone-labeled acoustic states, have been used to reduce the number of states needed to represent the set of speech units. Larger units, including *diphones, demisyllables, syllables, whole-words* and even *phrases*, have all been used to better characterize coarticulation between adjacent sounds. Acoustic segment units have also been investigated [Lee 1988].

- *Units with Linguistic Context Dependency*: Different ways of incorporting linguistic context in a speech subword unit, such as double context dependent phones (often known as *triphones*) and *generalized triphones*, have been proposed (e.g. [Lee 1989]). It has been shown that the recognition accuracy of a task can be increased when linguistic context dependency is properly incorporated to reduce the acoustic variability of the speech units being modeled. In fluent continuous speech it has also been shown that incorporation of interword units takes into account cross-word coarticulation and therefore provides more accurate modeling of speech units than simply using intraword context-dependent units. Word--dependent units have also been used to model poorly articulated speech sounds such as *function words* like *a, the, in, and*, etc. (e.g. [Lee 1989]).

For a given task, high recognition accuracy can be achieved only when the subword unit set contains context-dependent phones which maximally covers the vocabulary and the task language and when these phone units are adequately modeled using a large training set [Hon 1992]. However, the collection of a large amount of task-specific training data for every individual application is not practical. Task and *vocabulary independent* acoustic training and task-specific *vocabulary learning* (e.g. [Hon 1992, Lee *et al.* 1996]) are therefore important research topics. Task-independent modeling has also been applied to word spotting for training of acoustic models and rejection models [Rose and Hofstetter 1993, Sukkar and Lee 1996]. However, we do not yet know how to design a task-independent training database suitable for a wide range of vocabularies and

applications.

2.3 Acoustic Modeling of Speech Units

Training of subword unit models consists of estimating the model parameters from a training set of continuous speech utterances in which all of the relevant subword units are known to occur 'sufficiently' often. The way in which training is performed greatly affects the overall recognition system performance. A key issue in training is the size of the training set. Since infinite size training sets are impossible to obtain (and computationally unmanageable), we must use a finite size training set. This immediately implies that some subword units may not occur as often as others. Hence there is a tradeoff between using fewer subword units (where we get good coverage of individual units, but poor resolution of linguistic context), and more subword units (where we get poor coverage of the infrequently occurring units, but good resolution of linguistic context).

An alternative to using a large training set is to start with some initial set of subword unit models and adapt the models over time (with new training material, possibly derived from actual test utterances) to the task, the speaker and/or the environment. Such methods of adaptive training are usable for new speakers, tasks and environments, and provide an effective way of creating a good set of application-specific models from a more general set of models (which are speaker, environment, task, and context independent).

Speech patterns not only exhibit highly variable spectral properties but also show considerable temporal variation. There are not many modeling approaches that are both mathematically well-defined and computationally tractable, for modeling the speech signal. The most widely used and the most successful modeling approach to speech recognition is the use of hidden Markov models (HMMs). The reader is referred to a tutorial by Rabiner (1989) for an introduction to the HMM approach and its applications. Artificial neural network (ANN) approaches have also been used to provide an alternative modeling framework and a new computing paradigm [Bourlard and Wellekens 1992, Bourlard and Morgan 1994, Robinson 1994]. Almost all modern speech recognition systems use hidden Markov models and their extensions to model speech units. We will give a more detailed description of the HMM framework in the next section.

2.4 Lexical Modeling and Word Level Match

The second component of the word-level match module is the *lexicon* which provides a description of the words in the task vocabulary in terms of the basic set of subword units.

The lexicon used in most recognition systems is extracted from a standard dic-

tionary and each word in the vocabulary is represented by a single lexical entry (called a baseform) which is defined as a linear sequence of phone units. This lexical definition is basically *data-independent* because no speech or text data are used to derive the pronunciation. Based on this simplification, the lexical variability of a word in speech is characterized only indirectly through the set of sub-word models. To improve the lexical modeling capability, *data-dependent* approaches such as *multiple pronunciation* and *pronunciation networks* for individual words have been proposed (e.g. [Riley 1991, Bahl *et al.* 1993a]).

Among the issues in the creation of a suitable word lexicon is the baseform (or standard) pronunciation of each word as well as the number of alternative pronunciations provided for each word. The baseform pronunciation is the equivalent, in some sense, of a pronunciation guide to the word; the number of alternative pronunciations is a measure of word variability across different regional accents and talker population.

In continuous speech, the pronunciation of a word can change dramatically from that of the baseform, especially at word boundaries. It has been shown that multiple pronunciations or pronunciation networks can help deal with lexical variabilities more directly (e.g. [Riley 1991]).

Modeling lexical variability requires incorporation of language-specific phonological rules, the establishment of consistent acoustic-to-linguistic mapping rules (related to the selection and modeling of subword units), and the construction of word models. *Probabilistic word modeling*, which directly characterizes the lexical variability of words and phrases, is a promising research direction.

2.5 Language Modeling and Sentence Match

The sentence-level match module uses the constraints imposed by a grammar (or syntax) to determine the optimal sentence in the language. The grammar, consisting of a set of syntactic and semantic rules, is usually specified based on a set of task requirements. Although there have been proposed a number of different forms for the grammar (e.g. context-free grammar, *N*-gram word probabilities, word pair, etc.), the commonly used ones can all be represented as finite state networks (FSNs). In this manner it is relatively straightforward to integrate the grammar directly with the word-level match module.

The language models used in smaller, fixed-vocabulary tasks are usually specified manually in terms of deterministic finite state representations. For large vocabulary recognition tasks, stochastic *N*-grams such as bigram and trigram models have been extensively used (e.g. [Jelinek 1985]). Due to the sparse training data problem,

smoothing of the *N*-gram probabilities is generally required for cases with $N \geq 2$. *Class-dependent* bigrams and trigrams have also been proposed. To account for longer language constraints, tree language models have been proposed [Bahl *et al.* 1989]. The use of a *context-free language* in recognition [Ney 1991] is still limited mainly due to the increase in computation required to implement such grammars.

Advances in language modeling are needed to improve the efficiency and effectiveness of large vocabulary speech recognition tasks. Some of the advances will come from better stochastic language modeling. However the language models, obtained from a large body of domain-specific training data, often cannot be applied directly to a different task. *Adaptive language modeling*, which combines information in an existing language model and a small amount of application-specific text data, is an attractive approach to circumvent such difficulties.

2.6 Search and Decision Strategies

In addition to the use of hidden Markov models to model speech units, the other key contribution of speech research is the use of data structures for optimally decoding speech into text. In particular we use a finite state representation of all *knowledge sources*, including the grammar for word sequences, the network representation of lexical variability for words and phrases, as well as for morphemic, syllabic, and phonemic knowledge used to form fundamental linguistic units, and the use of hidden Markov models to map these linguistic units to speech units. Based on this type of data structure, most knowledge sources needed to perform speech recognition can be integrated into a finite network representation of hidden Markov acoustic states, with each state modeling the acoustic variability of each speech sound and all state transitions representing the link between different knowledge sources according to the hierarchical structure of the spoken language. As a result, speech recognition problems can be mapped to finding the most likely sequence of words through the task network such that the likelihood of the speech signal (or the corresponding acoustic feature vector sequence) is maximized. Decoding of such a network is accomplished efficiently through dynamic programming approach [Sakoe and Chiba 1978]. We give a detailed description of the DP search method in the next section.

3. Two Key Technologies for ASR

Two keys to the success of modern speech recognition systems are the use of hidden Markov modeling techniques to characterize and model the spectral and temporal variations of basic subword units (e.g. [Rabiner 1989]), and the use of dynamic pro-

gramming search techniques to perform network search (often referred to as *decoding*) to find the most likely sequence of words through a finite state network representation of a complex task (e.g. [Rabiner 1989]). We now give a brief description of each of these technologies.

3.1 Hidden Markov Modeling of Speech

The hidden Markov model is a statistical model that uses a finite number of states and the associated state transitions to jointly model the temporal and spectral variations of signals. It has been used extensively to model fundamental speech units in speech recognition because the HMM can adequately characterize both the temporal and spectral varying nature of the speech signal [Rabiner 1989, Rabiner and Juang 1993].

Although many variants exist, perhaps the simplest subword model is a left-to-right HMM with only *self* and *forward* transitions. Within each state of the model there is an observation density function which specifies the probability of a spectral vector. This observation density can either be a *discrete density* (implying the use of one or more codebooks to discretize the input spectral vector, e.g. [Lee 1989]), or a *continuous mixture density* (e.g. [Lee *et al.* 1990]), or a so-called *semi-continuous density* (e.g. [Huang and Jack 1989]) or a *tied-mixture density* (e.g. [Bellegarda and Nahamoo 1990]) which is a set of *common* continuous densities whose weights are chosen according to the model state. Tying can also be done at the HMM state level or at the state distribution level (e.g. [Hwang and Huang 1993, Young, Odell and Woodland 1994]). *Stochastic segment modeling* (e.g. [Lee, Soong and Juang 1988, Ostendorf and Roukos 1989, Deng 1993]), *dynamic system modeling* (e.g. [Digalakis, Rohlicek and Ostendorf 1993]), and *stochastic trajectory modeling* [Gong and Haton 1994] and *successive state splitting* [Takami and Sagayama 1992] have also been proposed to extend the HMM to handle intra-state, inter-state, and inter-sample correlations in a more precise manner. Some of the most often used acoustic modeling approaches include:

- *Maximum Likelihood (ML) Estimation of HMM*: Estimation of HMM parameters is usually accomplished in a batch mode using the ML approach based on the EM (estimation-maximization) algorithm (e.g. [Baum *et al.* 1970, Liporace 1982, Juang 1985]). Segmental ML approaches have also been extensively used (e.g. [Rabiner, Wilpon and Juang 1986]). Although ML estimation has good asymptotic properties, it often requires a large size training set to achieve reliable parameter estimation. Smoothing techniques, such as *deleted interpolation* [Jelinek and Mercer 1980] and *Bayesian smoothing* [Gauvain and Lee 1992], have been proposed to circumvent some of the problems associated with sparse training data.

- *Maximum Mutual Information (MMI) Estimation of HMM*: Instead of maximizing the likelihood of observing both the given acoustic data and the transcription, the MMI estimation procedure maximizes the mutual information between the given acoustic data and the corresponding transcription [Bahl *et al.* 1986, Normandin and Morgera 1991]. As opposed to ML estimation, which uses only class-specific data to train the classifier for the particular class, MMI estimation takes into account information from data in other classes due to the necessary inclusion of all class priors and conditional probabilities in the definition of mutual information.

- *Maximum A Posteriori (MAP) Estimation of HMM*:   Perhaps the ultimate way to train subword units is to adapt them to the task, to the speaking environment, and to the speaker. One way to accomplish adaptive training is through Bayesian learning in which an initial set of seed models (e.g. speaker-independent or SI models) are combined with the adaptation data to adjust the model parameters so that the resulting set of subword models matches the acoustic properties of the adaptation data. This can be accomplished by maximum a posteriori estimation of HMM parameters [Lee, Lin and Junag 1991, Gauvain and Lee 1992, Gauvain and Lee 1994] and has been successfully applied to HMM-based speaker and context adaptation of whole-word and subword models. On-line adaptation, which continuously adapts HMM parameters and hyperparameters, has also been developed (e.g. [Huo and Lee 1996]).

- *Minimum Classification Error (MCE) Estimation of HMM and ANN*: One new direction for speech recognition research is to design a recognizer that minimizes the error rate on task-specific training data. The problem here is that the error probability is not easily expressed in a close functional form because the true probability density function of the speech signal is not known. An alternative is to find a set of model parameters that minimizes the recognition error based on a given set of application-specific, training or cross-validation data [Juang and Katagiri 1992]. Each training utterance is first recognized and then used for both positive and negative learning by adjusting the model parameters of all competing classes in a systematic manner. For HMM-based recognizers, a family of *generalized probabilistic descent* (GPD) algorithms has been successfully applied to estimate model parameters based on the minimum classification error criterion (e.g. [Katagiri, Lee and Juang 1991, Chou, Juang and Lee 1992, Juang and Katagiri 1992, Su and Lee 1994]). The MCE/GPD approaches are also capable of maximizing the *separation* between models of speech units so that both discrimi-

nation and robustness of a recognizer can be simultaneously improved.

3.2 Dynamic Programming Structural Search

There are two basic search strategies, the *modular* and the *integrated* approaches, to find the most likely sentence that satisfies all the acoustic and linguistic constraints. In the integrated approach, the recognition decision is made by jointly considering all the knowledge sources. In principle, this strategy achieves the highest performance if all the knowledge sources can be completely characterized and fully integrated. Using the knowledge hierarchy in the linguistic structure of acoustics, lexicon, syntax and semantics, it is possible to approximate some of the above knowledge sources and compile them into a single finite state network composed of acoustic hidden Markov model states and grammar nodes and their connections [Levinson 1985]. Speech recognition is then solved by matching the input feature vector sequence to all the sequences of possible acoustic state sequences and finding the most likely sequence of words traversing the above knowledge network (e.g. [Bahl, Jelinek and Mercer 1983]). This is the commonly adopted search strategy in speech recognition today. However, there are a number of problems with the integrated approach. First, not all knowledge sources can be completely characterized and integrated. For example, supra-segmental information such as prosody and long-term language constraints such as trigram word probabilities cannot be easily cast into the finite state specification. Second, for large vocabulary tasks, the compiled network is often too large and therefore it becomes computationally intractable to find the best sentence.

On the other hand, for the modular approach shown in Figure 3, the recognized sentence can be obtained by performing unit matching, lexical matching, and syntactic and semantic analysis in a sequential manner. As long as the interface between adjacent decoding modules can be completely specified, each module can be designed and tested separately. Therefore collaborative research among different groups working on different components of the system can be carried out to improve the overall system performance. A majority of existing spoken language understanding and dialogue systems are designed collaboratively in this manner among speech and natural language researchers. In addition, modular approaches are usually more computationally tractable than integrated approaches. However one of the major limitations with the modular approach is that hard decisions are often made in each decoding stage without knowing the constraints imposed by the other knowledge sources. Decision errors are therefore likely to propagate from one decoding stage to the next and the accumulated errors are likely to cause search errors unless care is taken to minimize hard decision errors at every decoding or matching stage (e.g. retaining multiple hypotheses at each stage).
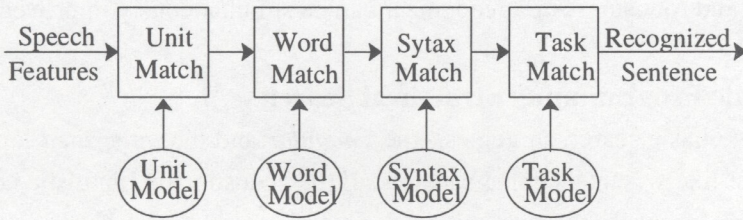
Figure 3 *Block diagram of a typical modular continuous speech recognizer*

Significant progress has been made in developing effective search algorithms in the last few years, including:

- *One-Pass Frame-Synchronous Beam Search*:   For the part of the knowledge that can be integrated into a finite-state network, the search problem is usually solved by finding the most likely path through the network. A full *breadth-first* search, such as the Viterbi algorithm (e.g.[Rabiner 1989]), can be very expensive in terms of processing time and storage requirements. To reduce computation, sub-optimal search strategies are commonly used. In the frame-synchronous beam search approach, only a small set of plausible partial (word) hypotheses within a beam are extended at any time instant. The resulting procedure is an approximate breadth-first search algorithm which requires the entire search space be maintained at all times during processing. Tree lexicon and phone look-ahead techniques have been implemented to reduce the number of plausible partial hypotheses in the beam search (e.g. [Ney *et al.* 1992]). The beam width, which determines the computation costs and possible pruning errors, is usually set experimentally and is both task-dependent and model-dependent.

- *Stack Decoding and A* Heuristic Search*:   Since the speech signal carries linguistic information in a somewhat localized manner, not all the linguistic events are active and need to be evaluated at the same instance of time. To take advantage of this property, a *best-first* search strategy can be used. The search is usually implemented using a stack (e.g. [Paul 1991, Soong and Huang 1991]) which maintains an ordered list of partial theories at every time instant. The best theory in the stack is extended first to a small list of word extensions according to the goodness of the acoustic and language-level matches. The extended theories are re-inserted into the stack. One advantage of this method is that long-term language models can be integrated naturally into the search. A* search strategies are usually used

to maintain the *admissibility* of search and to help limit the exponential growth of the stack size by invoking heuristic knowledge of the likely future theory extensions. The tree-trellis algorithm [Soong and Huang 1991] is an efficient way to achieve admissibility by maintaining all forward partial theories in the *forward trellis search* and recombining them with backward partial theories accumulated in the stack in the *backward tree search* to obtain a set of optimal theories. To reduce the number of word extensions, the use of a beam search, and an acoustic *fast match* for words and good acoustic and language models are essential.

- *Multi-Pass Decision Strategies*: As opposed to the traditional left-to-right, one-pass search strategies, multi-pass algorithms perform a search in a way that the first pass typically prepares partial theories and additional passes finalize the complete theory in a progressive manner. Multi-pass algorithms, such as the abovementioned tree-trellis algorithm, are usually designed to provide the *N*-best string hypotheses (e.g. [Schwartz and Chow 1990, Soong and Huang 1991]). To improve flexibility, simpler acoustic models can be used to produce a *segment lattice* [Zue *et al.* 1989], or a *phone lattice* [Ljolje and Riley 1992] in the first pass of a rough match. Lexical and language models can then be incorporated to generate a *word lattice*. Detailed models and detailed matches are then applied in later passes to combine partial theories into the recognized sentence. This family of techniques, sometimes referred to as *progressive search* [Murveit *et al.* 1993], is an efficient way to combine different level of knowledge sources in a systematic manner to improve speech recognition.

In the future, it seems reasonable to assume that a *hybrid search* strategy, which combines a modular search with a multi-pass decision, will be used extensively for large vocabulary recognition tasks. Good *delayed decision* strategies in each decoding stage are required to minimize errors caused by hard decisions. Approximate admissible fast matches are also needed to speed up decisions (e.g. [Bahl *et al.* 1993a, Kenny *et al.* 1993]). Multiple word and string hypothesization is also crucial for the integration of multiple and sometimes incompatible knowledge sources. The *N*-best search paradigm (e.g. [Schwartz *et al.* 1992]) is an ideal way for integrating multiple knowledge sources. It has been used for rescoring a preliminary set of candidate digit strings with higher-level constraints like a digit check-sum [Soong and Huang 1991], with detailed cross-word unit models and long-term language models [Schwartz *et al.* 1992], and with segmental neural nets [Zavaliagkos *et al.* 1994], etc. It has also been used to provide competing string hypotheses for discriminative training and for combining multiple acoustic models to reduce recognition errors [Chou, Lee and Juang 1993]. We expect to see more use of

the *N*-best search paradigm for incorporating high-level knowledge which cannot easily be integrated into the finite state representation for frame-based DP search. When combined with good multi-pass search (e.g. [Murveit *et al.* 1993]) and utterance verification (e.g. [Rahim, Lee and Juang 1995, Rose and Hofstetter 1993, Sukkar and Lee 1996]) strategies, it can effectively improve the flexibility and efficiency in designing large vocabulary spoken language systems.

4. Continuous Speech Recognition of Mandarin

In the following we illustrate, by example, how the pattern recognition approach we discussed above can be used to put together an integrated continuous speech recognition system for Mandarin. The algorithm described here was first developed for recognition of American English [Lee *et al.* 1990, 1992]. The same strategy was also applied to recognition of Spanish [Alvarez-Cercardillo *et al.* 1995] and Mandarin Chinese as shown here. Only the language-dependent part of the algorithm was modified to deal with the lexical and language constraints of different languages. Although the selection of an appropriate set of fundamental speech units for recognition is a crucial part, we assume only little knowledge about the phonology of a particular language in our studies. It is important to note that the examples shown here are meant to explore new research dimensions in Mandarin recognition. It is not our intention to get the best recognition results for any of the tasks, discussed here. Therefore no system parameter tuning was attempted in all of our experiments. Although the tonal property is a unique feature of Mandarin (as well as in some other Oriental languages), we do not address the issue in this paper. We only consider Mandarin *base syllables*, i.e. we do not distinguish syllables with only tone differences. The readers are referred to a number of studies for tone properties of Mandarin (e.g. [Tseng 1981]), tone recognition (e.g. [Wang and Lee 1994, Wang and Chen 1994]), and combined tone and syllable recognition (e.g. [Lin *et al.* 1993]).

For some historical and practical reasons, most of the existing Mandarin recognition systems developed in Taiwan and Mainland China are syllable based. Although modeling of sub-syllabic units has been attempted recently, most systems still perform syllable recognition in terms of a syllable lattice followed by lexical and language processing to determine the recognized sentence (e.g. [Lee *et al.* 1993, Chiang, Lin and Su 1996]). Because real-time system implementation has always been a major design factor in developing Mandarin recognition systems, this decoupled approach was largely adopted to avoid the search complexity implied in integrated approaches. Although some success has been reported for speaker-trained recognition, large-scale, speaker-

-independent Mandarin recognition has not been widely studied. Recall in our discussion above that due to error propagation from one stage to the next, decoupled approaches often degrade in performance more drastically than that in integrated approaches when inadequate intermediate search theories are saved for further processing. This is of major concern especially when testing in adverse conditions.

4.1 Baseline System

The baseline system we used for training and recognition is described in detail in [Lee *et al.* 1992] . Input speech, sampled at 8 kHz, was initially pre-emphasized $(1-0.95z^{-1})$ and grouped into frames of 240 samples with a shift of 80 samples. For each frame, a Hamming window was applied followed by a 10th order LPC analysis. A liftered 12-dimensional LPC-derived cepstral vector was then computed. The first and second time derivatives of the cepstrum were also computed. Besides the cepstral-based features, the log-scaled energy, normalized by the peak, and its first and second order time derivatives were also computed. Thus, each speech frame was represented by a vector of 39 features [Lee *et al.* 1992]. Speech recognition is performed using a frame-sychchronous beam search algorithm [Lee *et al.* 1992].

4.2 HKU93 - A Mandarin Database

The database we used for studying Mandarin continuous speech recognition is the HKU93 Putonghua Corpus made available to us by Department of Computer Science, University of Hong Kong [Zu, Li and Chan 1994]. The HKU93 corpus consists of a total of 20 native Putonghua speakers, 10 females and 10 males, each speaking: (1) all Putonghua syllables in all tones at least once, (2) 11 words of 2 to 4 syllables, (3) 16 digit strings of 4 to 7 digits, (4) 3 sentences of 7 rhymed syllables with /a/, /i/ and /u/ endings respectively, and (5) hundreds of sentences with verbalized punctuation from newspaper text. All speech recording were made in a quite room with a single National Cardioid Dynamic Microphone. Speech was digitized using a Sound Blaster 16 ASP A/D card plugged into a 486PC at 16-bit accuracy and with a sampling rate of 16KHz.

To simulate telephone speech recognition for our interest, we low-passed the speech signal and down-sampled it to 8KHz sampling rate. White noise was also added to produce a noisy version of the signal at about 35dB overall signal-to-noise ratio. Except for one of the speakers whose speech data were corrupted, we used all the data of the remaining 19 speakers for all the recognition experiments. Speech data from 16 speakers, 8 males and 8 females, were used for training and the other 3 speakers, 2 males and 1 female, were used for adaptation and testing.

4.3 Selection of Speech Units

Selection of fundamental speech units and acoustic modeling of such a set of speech units are two key steps in the design of a large vocabulary speech recognition system for any spoken language. In this paper, we study these two important issues for speaker independent continuous speech recognition of Mandarin.

4.3.1 Context-Independent (CI) Phone Set

The simplest way to obtain a set of task independent phone models is to choose the set of CI units, each of them modeling respectively the phoneme of the language. There is no context mismatch problem here. For American English, we adopted a set of 40 phonemes commonly used in the Bell Labs' Text-to-Speech grapheme-to-phoneme transcription rules [Lee *et al.* 1996]. In this study for Mandarin Chinese, we assume no knowledge of Mandarin phonology is available. Therefore we adopted the simplest CI phone set, namely the set of the 37 phonetic symbols, instead of using syllabic units or the corresponding sub-syllabic units as basic speech units, as commonly adopted in most Mandarin recognition systems. These 37 CI phone units correspond to the 37 essential Mandarin phone symbols, 22 syllable initials and 15 syllable medials and finals, commonly used in Taiwan for phonetic labeling of Chinese text. In addition to the above basic set, some 38 multi-phone syllable-final units are often used to replace the above set of 15 single-phone units. These units can be formed by concatenating the corresponding CI units according to a sub-syllabic lexicon. This gives a second set of 60 basic phone units consisting of these two types of sub-syllabic units commonly known in Taiwan as *shen-mu* and *yun-mu*. Since units in these two phone sets often occur enough times in a training set of reasonable size, they can be trained with most conventional acoustic modeling techniques.

4.3.2 Double-Context Phone Sets

To improve modeling accuracy, we then incorporated context-dependent (CD) units. Two unit selection approaches are studied. The first is the conventional method which selects only the triphone units that appear in the training set with enough occurrences. This strategy is usually more effective for task-dependent (TDEP) training which is often designed to handle test sentences that have a similar content coverage to that of the training sentences. If the context coverage of the test data is very different from that of the training data, it is likely that only a small portion of the phone units are actually used in recognition for the particular task. This is undesirable because only a small portion of the limited training data is used effectively and it often results in a poor recognition performance (e.g. [Lee *et al.* 1993]).

4.3.3 Single-Context Phone Sets

The second method is to select the complete set of right CD (RCD) units. Unlike triphone units that depend on both the right and the left context, the RCD unit only depend on the right context. In principle, there are a total of 37x38=1369 RCD units (including transition to the background silence) in Mandarin. However, since not all single-context CD phone units appear in the training set and not all units appear frequently enough, we use an occurrence threshold of 50 to limit the number of single-context units. This resulted in 718 right context-dependent (RCD) phone units (as opposed to the full set of 1406 units). To deal with phone units not represented by this RCD set we also supplement the right CD phone set with the set of CI units which bring the RCD set to a total of 756 units. This phone set can also be combined with other double-context unit sets to form larger phone sets that cover a wider range of context.

In a separate study [Lee *et al.* 1996], we have found that the RCD unit set is appropriate for task-independent (TIND) training which is designed to handle testing sentences that might have different context coverage from that in the training set.

4.4 Acoustic Modeling of Speech Units

Given the set of CI or CD units, we model each speech unit as a continuous density hidden Markov model (CDHMM). Except for the background silence unit, each subword unit is modeled by a 3-state left-to-right HMM with no state skip. Each state is characterized by a mixture Gaussian state observation density. Training is initially done with an iterative segmental ML algorithm (e.g. [Lee *et al.* 1992]) in which all utterances are first segmented into subword units. The Baum-Welch algorithm is then used to estimate the parameters of the mixture Gaussian densities (e.g. [Lee *et al.* 1992]) for all states of subword HMMs. Recognition is accomplished by a frame synchronous beam search algorithm [Lee *et al.* 1992] to determine the sequence of words (or phones) that maximizes the likelihood of the given utterance.

4.4.1 MCE-Based Phone Model Training

The conventional MCE training [Chou, Lee and Junag 1993] aiming at minimizing word recognition error in task-dependent training case, no longer provides the best model set because the target vocabulary is not fixed in testing. To alleviate this difficulty, we propose to replace the minimum word error objective with a new minimum phone error objective. This allows us to use the same algorithm; all we need to do is to first change the transcription of each utterance by simply replacing each word with its corresponding phone transcription through the use of the training lexicon. We then perform phone

recognition (instead of word recognition) when generating the *N*-best competing phone strings for each given utterance phone transcription (*N*=4 in all of our experiments). For a typical phone set of less than 1,000 units, this is more efficient than word recognition because in a typical TIND training set, there is a long list of distinct words (usually more than 5,000) which makes it a very slow process to obtain the *N*-best competing word strings for each utterance during MCE training.

The model parameters were estimated using the segmental GPD algorithm [Chou, Lee and Junag 1993]. The same algorithm can be used for both CI and CD model training. In this study we performed five iterations of GPD training for CI models and seven iterations of GPD training for CD models. CI and CD models of different sizes were created. For CI models, we have a maximum of 16 mixture components for each HMM state. For the single- and double-context CD models, we used a maximum of 8 mixture components per state.

4.4.2 Speaker Adaptive Phone Model Training

An alternative to using a large training set is to use some initial set of subword unit models and adapt them over time (with new training material, possibly derived from actual test utterances) to the speaker and speaking environment. In principle adaptation can be performed on lexical, syntactical and semantic models. We will focus our discussion only on adaptive training of subword acoustic models. Such methods of adaptive training are reasonable for new speakers, vocabularies, transducer or environments, and will be shown later to be an effective way of bootstrapping a good set of specific models from a more general set of models.

For adaptive training of subword unit models, we assume available a set of initial models, called *seed models*. The seed model can be a set of speaker-independent subword models or a set of gender-dependent subword models. Based on a small number of adaptation utterances, the adaptation algorithm attempts to combine the seed models with the adaptation data and generate a set of speaker adaptive (SA) models. In doing so, the dispersed seed models (e.g. the speaker independent models), which were designed to cover a wide range of speaking environments and a large number of speakers in the test population, are modified by the adaptation data (e.g. speaker-specific, application-specific utterances) so that a more focused set of models is created. The adaptive models are therefore useful in a more specific environment for a smaller number of speakers whose speech has similar acoustic characteristics to those of the adaptation data. In our HMM-based system, we used the *segmental MAP algorithm* [Gauvain and Lee 1992, 1994] to perform adaptive training. Given the acoustic data and the word

transcriptions of the adaptation utterances, the Viterbi algorithm is first used to segment the adaptation utterances into subword segments using the seed models. The MAP estimation algorithm is then used to obtain the parameters of the adapted subword models. In essence, the MAP estimate is a weighted sum of the prior parameters and the statistics of the adaptation data [Gauvain and Lee 1992]. The weights are functions of both the prior parameters and the adaptation data, and are recomputed in a nonlinear manner using the *expectation-maximization* (EM) algorithm.

4.5 Lexical and Language Modeling

Lexical modeling refers to modeling of lexical items such as syllables, words and phrases using models of fundamental speech units. For Mandarin recognition, we are mostly interested in modeling syllables. In this study, each Mandarin syllable is modeled as a concatenation of a sequence of phone models according to a syllable lexicon. By disregarding the tones associated with each syllable, there are a total of 484 syllables in the HKU93 corpus, including a basic set of 410 syllables plus a set of 74 syllables with *retroflexed* endings [Zu *et al*. 1994]. This set of 484 syllables is considered to be more difficult to recognize than the set of 408 base syllables commonly used in Taiwan (e.g. [Lee *et al*. 1993]). Word models can be constructed similarly by concatenating syllable models. Inter-syllable coarticulation can also be considered in word modeling.

Language modeling refers to modeling of the inter-dependency among the recognition units and their interactions. The recognition units could be phones, syllables, words or even phrases. In this study, we mainly address language modeling of phones and syllables. Both uniform grammars, which assume an unit follows any particular unit with equal probability, and bigram are employed.

4.6 Recognition Experiments

In the following we use the above three sets of phone units, namely the CI set of 37 phone units, the triphone set of 1330 units, and the RCD set of 756 units, and their corresponding phone models to perform both phone and syllable recognition of Mandarin speech. Three types of models are used; they are trained by the ML, MCE and MAP algorithms respectively. The number following the model type in all the following tables indicates the maximum number of mixture Gaussian components used to characterize the state observation densities. For example, MCE-8 represents the 8-mixture models trained with the MCE algorithm.

A total of 42,000 utterances from 16 speakers, about 60% of them isolated syllable utterances, were used to train all speaker independent models. On the other hand, about

8,000 utterances from 3 speakers were used for most of the testing. For generating speaker adaptive models, the first 300 utterances from each of the three testing speakers were used as the adaptation data. These utterances are therefore not used in obtaining the results in Tables 4 and 8 shown in the following.

Two types of language models were used. The first is the uniform grammar which is represented in the UG rows in the following tables of results. The second type is the bigram model (represented by BG in all tables) which assigns different probabilities according to the likelihood of two phones or two syllables appear together in the written text. The unigram perplexity in phone for the HKU93 training text data is about 27 (as opposed to 37 for the uniform grammar). This perplexity is reduced to about 11 when bigram is used. Compared with the English phone perplexity of 31 and 18 for unigram and bigram respectively, the bigram perplexity for Mandarin is much less. This is due to the fact that the syllable structure in Mandarin shows that consonants are usually followed by vowels which makes consonant-to-consonant or vowel-to-vowel transitions less likely. This fact will be reflected again when using bigram constraints to perform phone recognition. As for syllable language models, the unigram perplexity is 116 which is already a big reduction from the uniform grammar perplexity of 484. The perplexity is further reduced to 61 when the syllable bigram is imposed.

4.6.1 Phone Recognition

Phone recognition assumes that the recognized units are phones. In this study, there are 37 phones to be recognized. Both CI and CD models can be used to perform phone recognition of continuous speech. Phone recognition errors using models of the abovementioned three sets of CI and CD units are listed in Tables 1 to 3 respectively. It can be seen that using the BG grammar often produced much better results than using the UG grammar. MCE models are usually better than ML models. More complex models (with more mixture components) often outperform less complex models (with less mixture components). RCD models seem to do as well as the triphone models, although the RCD set encompasses fewer models (756 vs. 1330). The results obtained with the MCE-8 models shown in Table 3 are even slightly better than those obtained with the ML-8 models shown in Table 2. Finally, speaker adaptation produced more focused models (MAP-8) than the general ones (ML-8) and gives about 25% error reduction for phone recognition as shown in Table 4.

4.6.2 Syllable Recognition

Syllable recognition is difficult because the perplexity is high and the vocabulary is confusable. Using bigram, the syllable perplexity is 56, considerably higher than the

Table 1. *Phone error rates (%) using CI model sets.*

-	ML-16	MCE-16	Error Reduction
BG	24.1	19.8	17.8
UG	15.1	14.5	11.9

Table 2. *Phone error rates (%) using triphone model sets.*

-	ML-4	ML-8	Error Reduction
BG	16.2	14.6	11.1
UG	21.3	19.9	6.6

Table 3. *Phone error rates (%) using RCD model sets*

-	ML-8	MCE-8	Error Reduction
BG	15.9	11.8	24.5
UG	22.1	16.6	24.9

Table 4. *Phone error rates (%) using adaptive model sets*

-	ML-8	MAP-8	Error Reduction
BG	16.8	11.8	29.8
UG	23.6	18.1	23.3

phone bigram perplexity of 11. Furthermore, there are many confusable syllable groups in which syllables differ from each other only by one phone. Syllable recognition errors using models of the abovementioned three sets of CI and CD units are listed in Tables 5 to 7 respectively. It can be seen that the error comparison patterns for syllable recognition are similar to those for phone recognition. However, syllable recognition errors are much higher than those for phone recognition, reflecting the increased recognition difficulty. It is also noted that syllable recognition is much easier for continuous speech than for isolated syllables when bigram is used. This is shown in Table 8 in which we compare continuous syllable recognition with bigram, shown in the BG(CON) row, and isolated syllable recognition with the uniform grammar, shown in the UG(ISO) row. It is

also interesting to see that the improvement of MAP-8 over ML-8 is more significant for continuous syllable recognition than for isolated syllable recognition (45% vs. 16% error reduction). This is due to the fact that all the 300 utterances of adaptation data used to create MAP-8 models for each speaker are continuous syllable utterances. It is obvious from here that syllable properties for continuous speech is very different from those for isolated syllables. In order to have a better adaptation effectiveness and efficiency, both isolated syllables and continuous syllable utterances need to be used in speaker adaptation.

Table 5. *Isolated syllable error rates (%) using CI model sets*

ML-16	MCE-16	Error Reduction
34.5	26.5	23.2

Table 6. *Continuous syllable error rates (%) using triphone model sets*

-	ML-4	ML-8	Error Reduction
BG	25.0	23.3	6.8
UG	32.7	29.2	7.6

Table 7. *Continuous syllable error rates (%) using RCD model sets*

-	ML-8	MCE-8	Error Reduction
BG	23.0	20.1	12.6
UG	30.3	26.8	11.6

Table 8. *Syllable error rates (%) using adaptive model sets*

-	ML-8	MAP-8	Error Reduction
BG(CON)	19.3	10.6	45.3
UG(ISO)	30.9	25.8	16.5

4.6.3 Word Recognition

In addition to performing phone and syllable recognition on continuous speech, word recognition of Mandarin can also be performed just like what's been done in many

existing tasks of many different languages. It is easy to extend what we have done for both phone and syllable recognition to word recognition once the vocabulary and the grammar of the new task are defined. Since we do not have enough test data from any particular tasks, we only report some informal results here. One task is recognizing 1,078 names, from 2 to 7 syllables, spoken in isolation. We have achieved about 89% correct recognition using the GPD-trained RCD models. This is not as good as what we were getting at 93% accuracy for a similar task of recognizing 1,200 New Jersey town names [Lee *et al.* 1996].

4.6.4 Recognition Experiment Summary

Based on the experimental results, we have made the following observations, namely: (1) CD models outperform CI models in the same recognition test; (2) GPD models out-perform ML models in the same recognition test; (3) phone recognition outperforms syllable recognition because of the large number of confusable syllables in Mandarin; (4) speaker adaptive training using continuous utterances is more effective for recognizing continuous syllable utterances; and (5) the additional 74 retroflexed syllables gave more errors than the other basic syllables as indicated in our error analysis of the results. In summary, for speaker independent testing, the best continuous syllable recognition rate is about 80% while the best phone recognition rate is about 88% using GPD-trained RCD models. Except for some language-specific issues, we feel that Mandarin recognition is no diffferent from recognition of other languages. Algorithmic issues discussed above and system issues discussed in the next section are all common research issues.

5. ASR Systems and Applications

We now give a performance assessment of the ASR technology and briefly discuss how it can be used. We show first how ASR systems for isolated word recognition and con-tinuous speech recognition perform under ideal laboratory conditions for American English. We then discuss how the recognition sub-system can be integrated into real-world applications.

5.1 Laboratory System Performance

A summary of the performance of speech recognizers, based on laboratory evaluations, for the three technology areas (isolated words, connected words, fluent speech), and for different task applications, is shown in Table 9. (The reader should note that real world performance of most recognition systems is significantly worse than that of the labora-tory evaluations shown in Table 9. The measure of recognizer performance is the word

error rate (in percent) for a given vocabulary, task perplexity, and syntax (grammar).

For simple tasks like isolated digit recognition the word error rates are quite low both in SD (speaker dependent) mode (0%) and in SI (speaker independent) mode (0.1%). For an alpha-digit vocabulary, consisting of the spoken letters of the alphabet, the digits, and three command words, all spoken over dialed-up telephone lines, word error rates are 4.5% for the SD mode and 7.0% for the SI mode. Considering the confusability among spoken letters, these results are actually quite impressive for telephone bandwidth speech. For a more distinctive vocabulary of 1109 basic English words with more than half of the them monosyllabic words, the word error rate is 4.3% (SD) with a limit amount of training data. A similar accuracy of 4.7% (SI) is achieved for a vocabulary of 1218 town names using vocabulary-independent training.

Table 9. *Word error rates (%) for laboratory-evaluated ASR systems*

Technology	Task	Mode	Vocabulary	Error
Isolated Words	Words Equally Probable	SD	10 Digits	0
			39 Alphadigits	4.5
			1,109 Basic English	4.3
		SI	10 Digits	0.1
			39 Alphadigits	7.0
			1,218 Names	4.7
Connected Words	Digit Strings (Known Length)	SD	10 Digits	0.1
		SI	11 Digits	0.2
	Airline System (perplexity = 4)	SD	129 Airline Words	0.1
Fluent Speech	RM (word-pair) (perplesity = 60)	SI	991 Words	3.0
	ATIS (bigram) (perplexity = 25)	SI	1,800 Words	3.0
	NAB/WSJ (trigram) (perplexity = 145)	SI	20,000 Words	12.0

For connected word recognition, word error rates for known length digit strings are again quite low at 0.1% (SD) and 0.2% (SI). Similarly, for an airline reservations task, with a grammar whose perplexity (average word branching factor) is low (4), the word error rate in SD mode is 0.1%. For fluent speech recognition, results are based on DARPA funded research on three tasks; namely a ships database task (Naval Resource Management), an airline travel task (ATIS), and speech read from the *Wall Street Journal*. The vocabulary sizes and grammar perplexities of these three tasks are 991 words (perplexity 60), 1800 words (perplexity 25), and 20,000 words (perplexity 145), with laboratory evaluation word error rates of 3.0%, 3.0%, and 12.0%, respectively.

5.2 Speech Recognition Applications

Based on the task specific model, there is a broad range of applications of speech recognition both within telecommunications and in the business arena. Five broad application areas are:

- Telecommunications; providing information or access to data or services over telephone lines. Two of the most widely used applications include the AT&T Voice Recognition Call Processing (VRCP) system to automate operator-assisted call handling, and the NTT ANSER system for limited home banking services by voice.

- Office/desktop; providing recognition capability on the desktop including voice control of PC and workstation environments, voice interaction with PC/ workstation programs (e.g., voice commands as a supplement to the use of menus in word processors, calendar control, spreadsheets, etc.), voice control of telephony functionality (e.g., repertory dialing from PC files, voice access of messages stored in the PC, etc.), forms entry, dictation.

- Manufacturing and business; providing recognition capability to aid in the manufacturing processes, e.g., quality control monitoring on an assembly line; as an aid in handling packages for sorting (e.g., Federal Express shipments and US Postal Service) and delivery.

- Medical/legal; providing recognition capability for creating various reports and forms, e.g., radiology reports, wills, legal briefs, diagnostic reports, pathology analyses, etc. Such reports are generally highly constrained documents with highly technical jargon that is used repeatedly in each document.

- Other applications; including use of speech recognition in toys and games (voice interactions with game playing machines), and as aids for the handicapped (e.g.,

voice control of wheelchair functionality).

A key aspect in the success of voice recognition applications is how well the human-machine interface has been designed so that the recognition system is truly easy to use. The goal of the human factors design is to delight the user with the ease of use and the apparent simplicity of the task. The human factors enter through the judicious design and use of prompts, reprompts, voice repair, as well as in the mode (auditory, visual, tactile), timing, and content of feedback to the user. The most challenging part is to detect when the user is having problems and to provide the user with intelligent voice repairs and reprompts so that the user can be lead out of the difficulties without being annoyed. Careful system designs as well as algorithmic advances in utterance verification are needed to improve the flexibility and robustness of ASR systems.

6. Summary

We have briefly surveyed the present state of automatic speech recognition. We have witnessed, in the last several years, significant advances in almost all areas of speech recognition. Some of the technologies have already been incorporated into stand-alone products and telecommunication services. In the future, we expect that more recognition systems will be deployed as new problems emerge and novel technologies are developed. In addition to the fundamental technology and system design issues addressed above, we single out three key research challenges for bridging the gap between and a laboratory system and a real-world ASR application. They are: (1) robust speech recognition to improve the usability of a system in a wide variety of speaking conditions for a large population of speakers; (2) robust utterance verification to relax the rigid speaking format and to be able to extract relevant partial information in spontaneous speech and attach a recognition confidence to it; and (3) high performance speech recognition through adaptive system design to quickly meet changing tasks, speakers and speaking environments.

It is also important to know that the ASR module is only one of the components needed to realize a spoken language system. System design issues such as the transducer used to capture the speech signal, the channel used to transmit speech, and the speaking environment are all critical parts of a system. Human factors considerations such as interface design, ease of use, natural and intelligent prompts, error recovery and voice repair are also key research issues. Finally, the field of spoken language system research is still in its infancy. In order to advance from speech recognition to speech understanding, spoken dialogue and spoken language translation, collaborative effort among researchers from different areas is definitely needed.

7. Acknowledgement

References

Acero, A. and R. Stern, "Environmental Robustness in Automatic Speech Recognition," *Proc. IEEE ICASSP-90*, 1990, pp.849-852.

Alvarez-Cercardillo, J., C.-H. Lee and L. A. Hernandez-Gomez, "Acoustic Modeling of Context Dependent Units for Large Vocabulary Speech Recognition in Spanish," *Proc. EuroSpeech-95*, Madrid, Sept. 1995.

Bahl, L.R., F. Jelinek and R.L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Trans. Pattern Analysis, Machine Intelligence*, Vol. 5, 1983, pp. 179-190.

Bahl, L.R., P.F. Brown, P.V. de Souza and R.L. Mercer, "Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition," *Proc. IEEE ICASSP-86*, Tokyo, 1986, pp. 49-52.

Bahl L.R., P.F. Brown, P.V. de Souza and R.L. Mercer, "Tree-Based Language Model for Natural Language Speech Recognition," *IEEE Trans. Acous., Speech, Signal Proc.*, 1989, Vol. 37, pp. 1001-1008.

Bahl, L.R., S.V. de Gennaro, P.S. Gopalakrishnan and R.L. Mercer (1993a), "A Fast Approximate Acoustic Match for Large Vocabulary Speech Recognition," *IEEE Trans. Speech and Audio Proc.*, Vol. 1, No. 1, 1993, pp. 59-67.

Bahl, L.R., J.R. Bellegarda, P.V. de Sousa, P.S. Gopalakrishnan, D. Nahamoo and M.A. Picheny (1993b), "Multonic Markov Word Models for Large Vocabulary Continuous Speech Recognition," *IEEE Trans. Speech and Audio Processing*, Vol. 1, 1993, pp. 334-344.

Baum, L.E., T. Petrie, G. Soules and N. Weiss, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *Annal Math. Stat.*, Vol. 41, 1970, pp. 164-171.

Bellegarda, J.R. and D. Nahamoo, "Tied Mixture Continuous Parameter Modeling for Speech

Recognition," *IEEE Trans. Acoustics, Speech, Signal Proc.*, Vol. 38, 1990, pp. 2033-2045.

Bellegarda, J.R., P.V. de Sousa, A. Nadas, D. Nahamoo, M.A. Picheny and L.R. Bahl, "The Metamorphic Algorithm: A Speaker Mapping Approach to Data Augmentation," *IEEE Trans. Speech and Audio Proc.*, Vol. 2, 1994, pp. 413-420.

Biem, A., S. Katagiri and B.-H. Juang, "Discriminative Feature Extraction for Speech Recognition," *Proc. IEEE NN-SP Workshop*, 1993.

Bourlard, H. and C.J. Wellekens, "Links between Markov Models and Multi-Layer Perceptron," *IEEE Trans. Pattern Analysis, Machine Intelligence*, Vol. 12, 1992, pp.1167-1178.

Bourlard, H. and N. Morgan, *Connectionist Speech Recognition − A Hybrid Approach*, Kluwer Academic Publishers, 1994.

Chiang, T.-H., Y.-C. Lin and K.-Y. Su, "On Jointly Learning the Parameters in a Character-Synchronous Integrated Speech and Language Model," *IEEE Trans. Speech and Audio Proc.*, Vol. 4, No. 3, 1996, pp. 167-189.

Chou, W., B.-H. Juang and C.-H. Lee, "Segmental GPD Training of HMM Based Speech Recognizer," *Proc. IEEE ICASSP-92*, 1992, pp. I-473-476, San Francisco.

Chou, W., C.-H, Lee and B.-H. Juang, "Minimum Error Rate Training Based on the *N*-Best String Models," *Proc. IEEE ICASSP*, 1993, pp. II-652-655, Minneapolis.

Cox, S.J. and J.S. Bridle, "Unsupervised Speaker Adaptation by Probabilistic Fitting," *Proc. IEEE ICASSP-89*, 1989, pp. 294-297, Glasgow.

Davis, S.B. and P. Mermelstein, "Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acous., Speech, Signal Proc.*, Vol. 28, 1980, pp. 357-366.

Deng, L., "A Stochastic Model of Speech Incorporating Hierarchical Nonstationality, " *IEEE Trans. Speech and Audio Proc.*, Vol. 1, 1993, pp. 471-475.

Deng, L. and D. Sun. " A Statistical Approach to Automatic Speech Recognition Using the Atomic Speech Units Constructed from Overlapping Articulatory Features," *J. Acous. Soc. Am.*, Vol. 95, 1994, pp. 2702-2719.

Digalakis, V.V., J.R. Rohlicek and M. Ostendorf, " ML Estimation of a Stochastic Linear System with the EM Algorithm and Its Application to Speech Recognition," *IEEE Trans. Speech and Audio Proc.*, Vol. 1, 1993, pp. 431-442.

Digalakis, V.V., D. Rtischev and L.G. Nuemeyer, "Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures," *IEEE Trans. Speech and Audio Proc.*, Vol. 3, 1995, pp. 357-366.

Flanagan, J.L., *Speech Analysis, Synthesis and Perception*, 2nd edition, Springer-Verlag, 1972.

Fant, G.. *Speech Sounds and Features*, MIT Press, 1973.

Furui, S., "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum, " *IEEE Trans. Acous., Speech, Signal Proc.*, Vol. 34, 1986, pp. 52-59.

Furui, S., "Unsupervised Speaker Adaptation Method Based on Hierarchical Spectral Clustering," *Proc. IEEE ICASSP-89*, Glasgow, 1989, pp. 286-289.

Gales, M.J.F. and S.J. Young, "Parallel model combination for speech recognition in noise," *Technical Report*, CUED/F-INFENG/TR135, 1993.

Gauvain, J.-L. and C.-H. Lee, "Bayesian Learning for Hidden Markov Models With Gaussian Mixture State Observation Densities," *Speech Communication*, Vol. 11, Nos. 2-3, 1992, pp. 205-214.

Gauvain, J.-L. and C.-H. Lee, "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. Speech and Audio Proc.*, Vol. 2, No. 2, 1994, pp. 291-298.

Ghitza, O.,"Auditory Nerve Feedback as a Basis for Speech Processing," *Proc. IEEE ICASSP-88*, 1988, pp. 91-94.

Gong, Y. and J.-P. Haton, "Stochastic Trajectory Modeling for Speech Recognition," *Proc. IEEE ICASSP-94*, 1994, pp. 57-60.

Hattori, H. and S. Sagayama, "Vector Field Smoothing Principle for Speaker Adaptation," *Proc. ICSLP-92*, Banff, 1992, pp. 381-384.

Hon, H.-W. and K.-F. Lee, "Vocabulary Learning and Environmental Normalization in Vocabulary-Independent Speech Recognition," *Proc. IEEE ICASSP-92*, 1992, pp. I-485-488.

Hon, H.-W., "CMU Vocabulary-Independent Speech Recognition System," *Ph.D. Thesis*, School of Computer Science, Carnegie Mellon Univ., 1992.

Huang, X. and M.A. Jack, "Semi-continuous hidden Markov models for speech signal," *Computer, Speech and Language*, Vol. 3, 1989, pp. 239-251.

Huo, O. and C.-H. Lee, "On-line Adaptive Learning of the Continuous Density Hidden Markov Model Based on Approximate Recursive Bayes Estimate," *to appear in IEEE Trans. Audio and Speech Proc.*, 1996.

Hwang, M. and X. Huang, "Share-Distribution Hidden Markov Models for Speech Recognition," *IEEE Trans. Speech and Audio Proc.*, Vol. 1, 1993, pp. 414-420.

Jelinek, F. and R.L. Mercer, "Interpolated Estimation of Markov Source Parameters from Sparse Data," in *Pattern Recognition in Practice*, edited by E. S. Gelsema and L. N. Kanal,

North-Holland, 1980, pp. 381-397.

Jelinek, F., "The Development of an Experimental Discrete Dictation Recognizer," *Proc. IEEE*, Vol. 73,1985, pp.1616-1624.

Juang, B.-H., " Maximum-Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains," *AT & T Technical Journal*, Vol. 64, 1985.

Juang, B.-H., "Speech Recognition in Adverse Conditions," *Computer, Speech and Language*, Vol. 5, 1991, pp. 275-294.

Juang, B.-H. and S. Katagiri, "Discriminative Learning for Minimum Error Classification," *IEEE Trans. Signal Proc.*, Vol. 40, 1992, pp. 3043-3054.

Junqua, J.-C., H. Wakita and H. Hermansky, "Evaluation and Optimization of Perceptually-Based ASR Front-End," *IEEE Trans. Speech and Audio Proc.*, Vol. 1, 1993, pp. 39-48.

Katagiri, S., C.-H. Lee and B.-H. Juang, "New Discriminative Training Algorithms Based on the Generalized Probabilistic Descent Method," *Proc. IEEE NN-SP Workshop* 1991, pp.299-308.

Kenny, P. *et al.*, " A*-Admissible Heuristics for Rapid Lexical Access," *IEEE Trans. Speech and Audio*, Vol. 1, 1993, pp. 49-58.

Lee, C.-H., F.K. Soong and B.-H. Juang, "A Segment Model Based Approach to Speech Recognition," *Proc. IEEE ICASSP-88*, 1988, pp. 501-504.

Lee, C.-H., L.R. Rabiner, R. Pieraccini and J.G. Wilpon, " Acoustic modeling for large vocabulary speech recognition, " *Computer Speech and Language*, Vol. 4, No. 2, 1990, pp. 127-165 .

Lee, C.-H., C.-H. Lin and B.-H. Juang, "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models," *IEEE Trans. Acous., Speech, Signal Proc.*, Vol. 39, No. 4, 1991, pp. 806-814.

Lee, C.-H., E. Giachin, L.R. Rabiner, R. Pieraccini and A.E. Rosenberg, "Improved Acoustic Modeling for Large Vocabulary Continuous Speech Recognition," *Computer, Speech & Language*, Vol. 6, No. 2, 1992, pp. 103-127.

Lee, C.-H., J.-L. Gauvain, R. Pieraccini and L. R. Rabiner, "Large Vocabulary Speech Recognition Using Subword Units," *Speech Communication*, Vol. 13, Nos. 3-4,1993, pp. 263-280.

Lee, C.-H., B.-H. Juang, W. Chou and J. J. Molina-Perez, " A Study on Task-Independent Subword Selection and Modeling for Speech Recognition," *Proc. ICSLP-96*, Philadelphia, Oct. 1996.

Lee, K.-F., *Automatic Speech Recognition – The Development of the SPHINX-System*, Kluwer Academic Publishers, Boston, 1989.

Lee, L.-S. *et al.*, "Golden Mandarin (I) - A Real-Time Mandarin Speech Dictation Machine for Chinese with a Very Large Vocabulary," *IEEE Trans. Speech and Audio Proc.*, Vol. 1, No. 2, No. 2, 1994, pp. 158-179.

Leggetter, C.J. and P.C. Woodland, "Speaker Adaptation of Continuous Density HMMs Using Linear Regression," *Proc. ICSLP-94*, 1994.

Levinson, S.E., " Structural Methods in Automatic Speech Recognition," *Proc. IEEE*, Vol. 73, 1985, pp. 1625-1650.

Lin, C.-H. Lin *et al.*, "A New Framework for Recognition of Mandarin Syllables with Tones Using Sub-Syllabic Units," *Proc. IEEE ICASSP-93*, Vol. II, 1993, pp. 227-230, Minneapolis.

Ljolje, A. and M.D. Riley, "Optimal Speech Recognition Using Phone Recognition and Lexical Access, " *Proc. ICSLP-92*, 1992, pp. 313-316.

Liporace, L.R., "Maximum Likelihood Estimation for Multivariate Observations of Markov Sources," *IEEE Trans. Information Theory*, Vol. 28, 1982, pp. 729-734.

Liu, F.-H., "A. Acero and R.M. Stern, Efficient Joint Compensation of Speech for the Effect of Additive Noise and Linear Filtering," *Proc. IEEE ICASSP-92*, 1992, pp. I-257-260.

Merhav, N. and C.-H. Lee, "A Minimax Classification Approach with Application to Robust Speech Recognition," *IEEE Trans. Speech and Audio*, Vol. 1, No. 1,1993, pp. 90-100.

Murveit, H., J. Butzberger, V. Digalakis and M. Weintraub, "Large-Vocabulary Dictation Using SRI's DECIPHER Speech Recognition System: Progressive Search Techniques," *Proc. IEEE ICASSP*, 1993, pp. II-319-322.

Ney, H., "Dynamic Programming Parsing for Context-Free Grammar in Continuous Speech Recognition," *IEEE Trans. Signal Proc.*, 1991, Vol. 39, pp. 336-340.

Ney, H., R. Haeb-Umbach, B.-H. Tran and M. Oerder, "Improvement in Beam Search for 10,000-Word Continuous Speech Recognition, " *Proc. IEEE ICASSP-92*, 1992, pp. I-9-12.

Normandin, Y. and D. Morgera, " An Improved MMIE Training Algorithm for Speaker-Independent Small Vocabulary, Continuous Speech Recognition," *Proc. IEEE ICASSP-91*, 1991, pp. 537-540.

Ostendorf, M. and S. Roukos, "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition, " *IEEE Trans. Acous., Speech, Signal Proc.*, Vol. 37, 1989, pp. 1857-1869.

Paul, D.B., "Algorithm for an Optimal A* Search and Linearizing the Search in the Stack Decoder," *Proc. IEEE ICASSP-91*, 1991, pp. 693-696.

Parthasarathy, S. and C.-H. Coker, "On Automatic Estimation of Articulatory Parameters in a

Text-to-Speech System," *Computer, Speech and Language*, 1992, Vol. 6, pp. 37-75.

Rabiner, L.R., J.G. Wilpon and B.-H. Juang, "A Segmental *K*-Means Training Procedure for Connected Word Recognition," *AT & T Tech. Journal*, Vol. 65, 1986, pp. 21-31.

Rabiner, L.R., " A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, Vol. 77, 1989, pp. 257-286.

Rabiner, L.R. and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

Rahim, M. and B.-H. Juang, "Signal Bias Removal for Robust Telephone Speech Recognition in Adverse Environments," *Proc. IEEE ICASSP-94*, 1994, pp. 445-448, Adelaide.

Rahim, M., C.-H. Lee and B.-H. Juang, "Robust Utterance Verification for Connected Digit Recognition," *IEEE ICASSP-95*, 1995, pp. 285-288, Detroit.

Rahim, M. and C.-H. Lee, "Simultaneous Feature and HMM Using String-Based Minimum Classification Error Training," *Proc. ICSLP-96*, Philadelphia, Oct. 1996.

Riley, M.D., "A Statistical Model for Generating Pronunciation Networks," *Proc. IEEE ICASSP-91*, Vol. 2, 1991, pp. 737-740.

Robinson, A., "An Application of Recurrent Nets to Phone Probability Estimation," *IEEE Trans. Neural Networks*, Vol. 5, 1994, pp. 298-305.

Rohlicek, J.R., "Word Spotting," in *Modern Methods of Speech Processing*, edited by R. Ramachandran and R. Mammone, Kluwer Academic Publishers, 1995.

Rose, R.C. and E.M. Hofstetter, "Task-Independent Wordspotting Using Decision Tree Based Allophone Clustering," *Proc. IEEE ICASSP-93*, 1993, pp. II-467-470.

Rose, R.C., E.M. Hofstetter and D. A. Reynold, "Integrated Models of Speech and Background with Application to Speaker Identification in Noise," *IEEE Trans. Speech and Audio*, Vol. 2, 1994, pp. 245-257.

Sakoe, H. and S. Chiba, "Dynamic Programming Optimization for Spoken Word Recognition," *IEEE Trans. Acous., Speech, Signal Proc.*, Vol. 26, 1978, pp. 52-59.

Sankar, A. and C.-H. Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition," *IEEE Trans. Speech and Audio Proc.*,1996, pp.190-202, Vol. 4, No. 3.

Schwartz, R., Y.-L. Chow and F. Kubala, "Rapid Speaker Adaptation Using a Probabilistic Spectral Mapping," *Proc. IEEE ICASSP*, 1987, pp. 633-636.

Schwartz, R., and Y.-L. Chow, " The *N*-Best Algorithm: An Efficient and Exact Procedure for Finding the *N* Most Likely Sentence Hypotheses," *Proc. IEEE ICASSP-90*, 1990, pp. 81-84.

Schwartz, S., Austin, F. Kubala, J. Makhoul, L. Nguyen and P. Placeway, "New Uses for the *N*-Best Sentence Hypotheses within The BBN BYBLOS Continuous Speech Recognition System," *Proc. IEEE ICASSP-92*, 1992, pp. I-1-4.

Seneff, S., "A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing," *J. Phonetics*, Vol. 16, 1988, pp. 55-76.

Soong, F.K. and E.F. Huang, " A Tree-Trellis Based Fast Search for Finding the *N*-Best Sentence Hypotheses in Continuous Speech Recognition," *Proc. IEEE ICASSP-91*, 1991, pp. 703-706.

Su, K.-Y. and C.-H. Lee, "Speech Recognition using Weighted HMM and Subspace Projection Approaches," *IEEE Trans. on Speech and Audio Proc.*, Vol. 2, No.1, 1994, Jan, pp. 69-79 .

Sukkar, R. and C.-H. Lee, "Vocabulary-Independent Discriminatively Trained Method for Rejection of Non-Keywords in Subword Based Speech Recognition," *to appear in IEEE Trans. Speech and Audio Proc.*, 1996.

Takami, J. and S. Sagayama, "A Successive State Splitting Algorithm for Efficient Allophone Modeling, " *Proc. IEEE ICASSP-92*, 1992, pp. I-573-576.

Varga, A.P. and R.K. Moore, "Hidden Markov Model Decomposition of Speech and Noise," *Proc. IEEE ICASSP-90*, 1990, pp. 845-848.

Wang, H.-M. and L.-S. Lee, "Tone Recognition for Continuous Mandarin Speech with Limited Training Data Using Selected Context-Dependent Hidden Markov Models," *Jour. Chinese Institute of Engineers*, Vol. 17, No.6, 1994, pp. 775-784.

Wang, Y.-R. and S.-H. Chen, "Tone Recognition of Continuous Mandarin Speech Assisted with Prosodic Information," *J. Acoust. Soc. Am.* Vol. 96, No. 5, 1994, pp. 2637-2645.

Wilpon, J. G., L. R. Rabiner, C.-H. Lee and E. R. Goldman, "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models," *IEEE Trans. Acous., Speech, Signal Proc.*, Vol. 38, No. 11, 1990, pp. 1870-1878.

Young, S.J., J.J. Odell and P.C. Woodland, "Tree-Based State Tying for High Accuracy Acoustic Modeling," *Proc. ARPA Human Language Technology Workshop*, Princeton, 1994.

Zavaliagkos, G., Y. Zhao, R. Schwartz and J. Makhoul, "A Hybrid Segmental Neural Net/Hidden Markov Model System for Continuous Speech Recognition," *IEEE Trans. Speech and Audio*, Vol. 2, 1994, pp. 151-160.

Zhao, Y., " A New Speaker Adaptation Technique Using Very Short Calibration Speech, " *Proc. IEEE ICASSP-93*, 1993, pp. II-592-595.

Zu, Y. Q., W.X. Li and C. Chan, "HKU93 - A Putonghua Corpus, Department of Computer Science," University of Hong Kong, 1994.

Zue, V., J. Glass, M. Phillips and S. Seneff, "The MIT Summit Speech Recognition System : A Progress Report," *Proc. DARPA Speech and Natural Language Workshop*, 1989, pp. 179-189.