AN OVERVIEW ON SPOKEN LANGUAGE PROCESSING

Chin-Hui Lee

Multimedia Communications Research Lab

Bell Laboratories, Lucent Technologies

Murray Hill, NJ 07974, USA

Tel: +1-908-582-5226

Fax: +1-908-582-7308

email: chl@research.bell-labs.com

Abstract

We review the science and technology of spoken language processing. We first discuss the most successful approach to spoken language processing, namely the pattern recognition approach. We illustrate how sophisticate statistical modeling techniques can be used together with a large amount of spoken and written examples to design models for high performance spoken language systems. We then point out the current capabilities and limitations of spoken language systems. Finally we discuss research challenges to enhance the capabilities and reduce the limitations.

1 Introduction

In the last two decades advances in *automatic speech recognition* (ASR) *natural language processing* (NLP) have triggered the development of a number of *spoken language system* 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, restricted-domain speech translation, and spoken dialogue system. With the introduction of *internet*, *intranet* and *world wide web* (WWW), we expect to see speech input/output (I/O) capabilities, in the user's own language, be incorporated into some of the existing web interfaces to improve *human computer interaction* (HCI) [83]. Furthermore, with the emergence of *computer telephony integration* (CTI) [16], we also anticipate to have available intelligent speech interfaces serving as voice agents to provide interactive problem solving capabilities over the world-wide communication and computer networks.

Due to the sophistication of computing and communications systems, there is an increasing demand for such systems to be equipped with intelligent multimedia use interfaces. Speech, by far, is the most direct and natural means for human beings to communicate with machines. Because of the human involvement in the communication chain, spoken language processing has emerged as a new exciting research field. It encompasses many vastly different key areas, including acoustics and transducers, signal processing, wired and wireless transmission, array processing, audio and visual perception, speech/image coding, recognition and synthesis, natural language understanding and generation, heuristic search and problem solving, information retrieval, knowledge representation, multimedia presentation, database management and design, human factors, etc. Inspired by its inter-discipline nature, we have witnessed collaboration among researchers in some of the abovementioned areas. New directions are constantly being pursued and new advances are regularly being made. For a sample of the progress, interested readers are referred to the recent publications of the Proceedings of the annual International Conference on Acoustics, Speech and Signal Processing (ICASSP), the biannual European Conference on Speech Communication and Technology (EuroSpeech), and the biannual International Conference on Spoken Language Processing (ICSLP).

Much of the recent effort in spoken language processing has been stimulated by the Advanced Research Project Agency (ARPA) of the United States, formerly known as D(efense)ARPA, which has funded research, under the human language technology (HLT) and spoken language system (SLS) programs, on three recent language recognition and underatnding projects, namely the Naval Resource Management (RM) task, the Air Travel Information System (ATIS) and the North American Business (NAB, previously known as the Wall Street Journal or WSJ) task. In Europe, many countries and research groups participated in the SUNDIAL project [54] under the Esprit program to jointly develop systems that can understand several European languages in multiple domains (including Eurorail train reservation). In Japan, spoken dialogue understanding and generation research was carried as one of the priority scientific areas from 1993 to 1996, funded by the Ministry of Education, Science, Sports and Culture [17]. It involved more 30 universities conducting studies in four key areas in spoken dialogue processing, namely speech recognition and synthesis, language analysis and generation, understanding and presentation of conceptual information, and dialogue modeling. It is clear that there is a world-wide interest in establishing natural multi-lingual, human-machine communication interfaces.

It is also clear that much of the advances was made by a collaborative community, in which responsibilities, such as collecting large speech and text corpora, defining common task, developing research tools, building research infrastructure, establishing common evaluation metrics, and educating the community, the funding agencies and general public, are shared among participating groups. As a result, we are enjoying a steady progress in the spoken language processing technology.

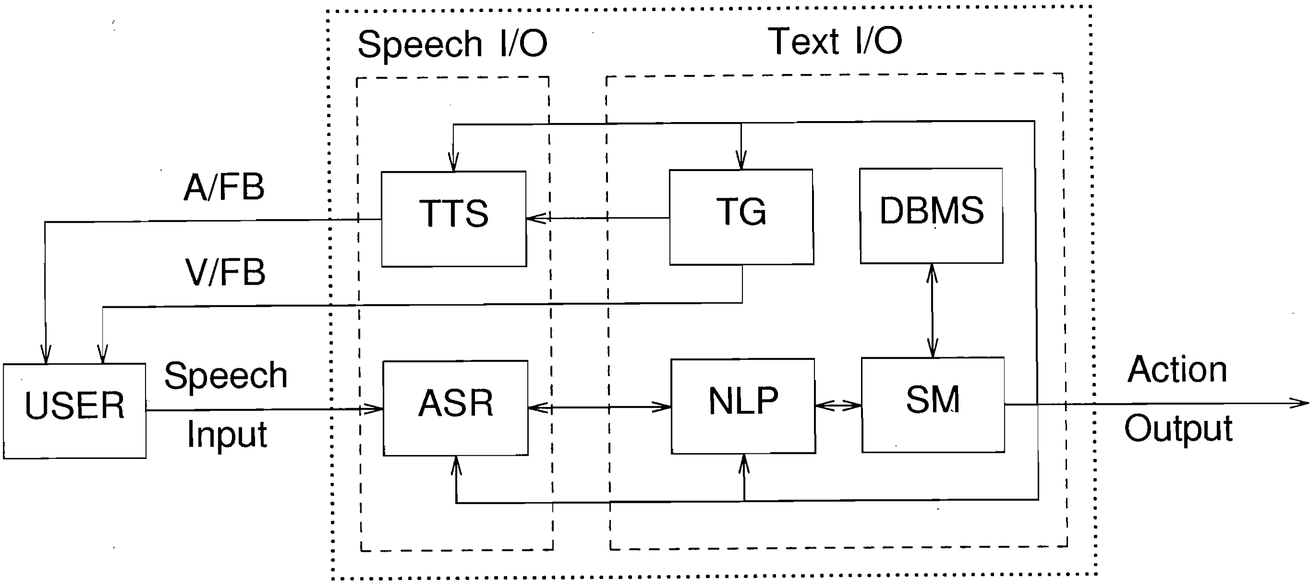

Figure 1: Block diagram of a typical spoken language system.

A block diagram of a typical spoken language system is shown is Figure 1. The USER module is a model of the user that produces speech input to the spoken language system shown in the dotted box. It also takes the information provided by the audio/visual feedback modules of the system. The spoken language system consists of two parts, namely the speech I/O component shown in the left dashed box and the text I/O components shown in the right dashed box. The automatic speech recognition module takes the speech input from the user and generates a preliminary list of recognized words, phrases and sentences. The natural language processing module analyzes the recognized partial hypotheses and produces a set of meaningful candidates for the system manager (SM) module. The SM module evaluates the current set of input candidates, communicates with the database management system (DBMS), prepares the output semantic action, and updates the discourse information which describes the current state of the dialogue session. This information can also be used as additional constraints to help reduce the computation requirement and to improve the output quality of both the ASR and the NLP modules. The text generation (TG) module extracts useful information provided by the SM module and generates a compact representation of the current state of the system so that a set of text, tables and graphics can be displayed as visual feedback (V/FB) information for the user. In some cases, the amount of materials to be communicated to the user is overwhelming. Information retrieval (IR) and systematic organization of the feedback information are required. A simple dialogue, via display and/or spoken interface, can even be established between the user and the system to properly communicate the extracted information to the user. While the text-to-speech (TTS) module generates a compact set of speech outputs to be played back as audio feedback (A/FB) for the user. Since the choices of the TTS and TG outputs depend on the availability and the quality of both the A/FB and the V/FB modules, they should be designed together to help enhance the human-machine interface of the system.

It is clear that a number of knowledge sources, including acoustic models of fundamental speech units, lexical models of words and phrases, syntactic models for word sequences, semantic models for meanings, dialogue models for monitoring the states of a dialogue session, are needed in the design of a spoken dialogue system. Many data-driven approaches, including the *connectionist* approaches based on *artificial neural networks* (ANN's) and *decision tree* approaches based on the *classification and regression tree* (CART) framework, can all be used to model the knowledge sources. It is also important to learn from realistic examples through data collection.

The rest of the paper is organized as follows. The two key ASR and NLP modules are discussed in details in Sections 2 and 3 respectively. Issues related to integration of ASR and NLP are addressed in Section 4. Since statistical pattern recognition paradigm is the most successful approach to speech recognition, a similar formulation for speech understanding is given in Section 5. The state-of-art technology is also reviewed. A case study of designing a real-world car reservations system in attempt to move spoken language system out of research laboratories is illustrated in Section 6. Some of the research issues in spoken language processing are also addressed. Finally we summarize our discussion in Section 7.

2 Automatic Speech Recognition

In this Section, we discuss in detail how speech recognition is implemented. We describe each building block, how blocks are put together and what are the research issues involved in speech recognition. This serves as an illustration how natural language processing problems can be solved and implementation. Although the system components are somewhat different and the research issues bear different dimensions, we believe the methodology adopted to solve speech and language problems are rather similar in nature.

The approach that is conventionally taken to speech recognition is basically a statistical pattern recognition approach. A block diagram of a typical subword-based continuous speech recognition system 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 the vocabulary words to determine which words were most likely spoken. The sentence-level match module uses a language model to determine the most likely sequence of words. Search and recognition decisions are made by considering all likely word sequences and choosing the one with the best acoustic matching score as the recognized sentence. Stochastic learning techniques play a crucial role both in the extraction of spectral features and in the design of the acoustic models for the fundamental speech units, the lexical models for the word and phrase units, syntactical models for the grammar and semantic models for the task constraints.

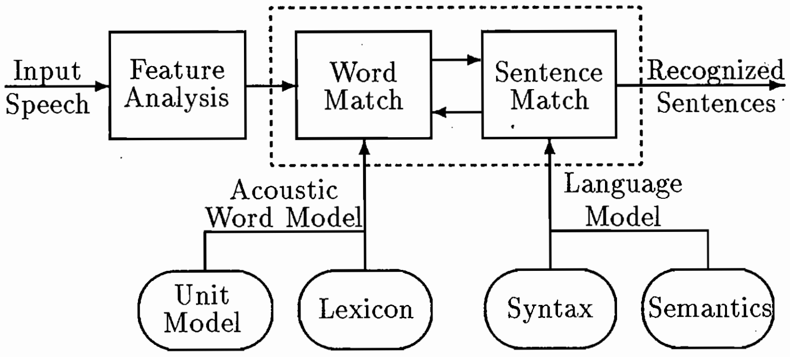

Figure 2: Block diagram of a typical integrated continuous speech recognizer.

Two keys to the success of modern speech recognition systems are the use of statistical modeling techniques (e.g. hidden Markov models, or HMM's) to characterize the basic subword units (e.g. [61]) and the use of dynamic programming techniques to search for the most likely sequence of words [8] of a knowledge network representing the recognition task. Since all the knowledge sources required to represent acoustics, morphology, lexicon, syntax and semantics are modeled by a finite state directed graph, this allows us to have an integrated knowledge network by embedding the multiple phone models into each lexical entry, and by embedding the multiple lexical entries into each word, and finally by embedding the word models into each sentence. The subword models are trained based on the network representation of the orthographic transcription of the training sentences. Recognition is then accomplished by finding the most likely (least costly) path in the network which implies the recognized word string. Each of the system component is described briefly in the following. For a survey of ASR research issues, the readers is referred to a recent article [43].

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.

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 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). Artificial neural network (ANN) approaches have also been used to provide an alternative modeling framework and a new computing paradigm. Almost all modern speech recognition systems use hidden Markov models and their extensions to model speech units.

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 dictionary 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.

Among the issues in the creation of a suitable word lexicon is the baseform (or standard) 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 units include diphones, demisyllables, syllables, whole-words and even phrases.

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. 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* are therefore important research topics (e.g. [44]).

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 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.

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. 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.

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. We will discuss the important issue of language modeling in Section 3.

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. We give a detailed description of the DP search method in Section 4.

3 Natural Language Processing

Similar to ASR, many NLP problems can often be formulated with the pattern matching paradigm as long as the language-related knowledge sources can be defined and modeled. However unlike ASR which is well defined as a problem of converting spoken utterances into a sequence of words, there are many more dimensions to be addressed in natural language processing. Depending of the task requirements, different NLP problems can be defined accordingly. The reader is referred to [30] in this Proceedings for a review. In this Section, We will focus our discussion mostly on stochastic language modeling. Issues related to spoken language understanding and dialogue processing will be addressed in Section 5.

Although stochastic modeling techniques have been used extensively in modeling speech units, the approach to modeling linguistic units, on the other hand, is largely centered around the rule-based, expert system paradigm. Rules needed for understanding natural language, including lexical, syntactic, semantic and discourse analysis, are usually specified manually by human experts based on task constraints. Knowledge representation, knowledge acquisition and task portability are three key issues in the design of rule-based language systems. When spoken language is considered as a means of natural language input, it creates additional processing demands on the system design. First, as opposed to the processing of text input, the speech input needs to be recognized by the system and converted to some form of text units for further processing. Therefore a good speech recognition "front-end" is absolutely required in a spoken language system. Second, since the recognition process is error-prone, the recognized text is often *ill-formed*, and the rules which are designed based only on text data may not be able to cope with the erroneous sentences produced by the speech recognition "front-end". Rule modification based on spoken data is therefore needed. Third, the grammars for spoken languages are different from those for written languages. Non-linguistic speech events, such as um's and ah's, false starts, disfluency and hesitation, etc., needed to be detected from the spoken input and identified along with all the other linguistically-defined speech events. Therefore, *spoken language grammar* modeling is an important research topic for designing good spoken language systems. Last but not the least, out-of-vocabulary spoken events, such as new words or ill-defined sentences, are difficult to identify which causes the number of falsely detected words to increase.

In the past several years, *corpus-based* language modeling approaches have emerged. Similar to corpus-based acoustic modeling which is used in most speech recognition systems, corpus-based language modeling requires a large body of labeled text data for training language models or deriving linguistic rules. A large set of test and/or cross-validation data to evaluate the performance of the modeling techniques is also needed. In contrast to speech which is a continuous signal, the text information is usually realized as a discrete event. There is no problem identifying fundamental text units, such as alphabets and words (as long as they are not mis-spelled). The properties associated with a fundamental text unit are usually represented as *attributes*, such as parts of speech, which are discrete in nature. Therefore, the most widely-used statistical modeling technique is to compute the N-gram probabilities of text units, including N-grams of alphabets, morphemes, syllables, words, classes of words, parts of speech, and semantic attributes.

The language models used in smaller, fixed-vocabulary tasks are usually specified manually in terms of deterministic finite state representations. *Class-dependent bigrams* and *trigrams* have also been proposed. To account for longer language constraints, tree language models have been proposed [3]. The use of a *context-free grammar* in recognition is still limited mainly due to the increase in computation and the difficulty in stochastic modeling. Only small context-free languages have been studied and used ([51]). A finite state approximation of a restricted-domain context-free language has also been evaluated for speech translation (e.g. [62]). Although, the use of word-based context-free grammars is limited, an LR parser has been successfully used to implement context-free phoneme look-ahead rules in a Japanese speech translation system (e.g. [50]).

3.1 Static Modeling of Linguistic Units

The most popular technique for characterizing discrete events is by counting their relative frequencies of occurrences in the training data. This results in the maximum likelihood estimate of the *unigram* probabilities [25]. By extending the same notion to counting of sequence of N consecutive discrete events, we have the maximum likelihood estimate of the N-gram probability of discrete events. For large vocabulary ASR tasks, stochastic N-grams such as bigram and trigram have been extensively used [33]. Due to the sparse training data problem, smoothing of the N-gram probabilities is often required for cases with $N \geq 2$. Another problem with maximum likelihood N-gram estimation is that many events which are not observed in training data often appear in testing. There using a null-probability as the estimate (MLE) for such events is not satisfactory. This null-probability problem is similar to the zero cell problem in the estimation of discrete HMM state distributions. Several smoothing techniques, including the backoff approach [37] the modified zero frequency technique (e.g. [57]), the add-one technique, and *class-dependent* N-grams, have been proposed to deal with distribution degeneracy problem and to improve the robustness and generalization capability of the N-gram models.

3.2 Modeling of Underlying Linguistic Structure

Words and sequence of words are observed events in written language processing just like speech signal is the observed event in spoken language processing. For ASR, the HMM framework has been successfully applied to characterize the unobserved (hidden) events, such as words, which are embedded in spoken language. Beyond words, there are other important unobserved events in spoken and written language processing, including classes of words such as parts of speech, word attributes such as meanings of words, structure of words such as grammar and the implied set of production rules associated with a sentence.

Stochastic modeling techniques have been applied to characterize such hidden linguistic structures. The first example is tagging a sequence of words with a sequence of parts of speech [14]. N-gram probabilities of observing sequence of parts of speech and the corresponding lexical probabilities are estimated from a labeled training corpus and dynamic programming search is used to determine the sequence of tags that maximizes the probabilities of observing the word and tag pairs. When the parts of speech labels are replaced with semantic attributes, we have a model of meanings expression as a sequence of semantic attributes and their interactions ([55]). The semantic attribute, called *concept*, are hidden and needed to be decoded from the observed sequence of words. The hidden Markov modeling framework can now be applied directly. N-gram probabilities are again used to model the concept sequence and the concept-specific language model in a semantic state ([55]).

In formal language characterization words are considered as *terminal* symbols which are directly observed in most written languages (although in some languages, such as Chinese and Japanese, words are not directly observed and need to be segmented because no space symbols are used as word boundary markers). *Non-terminal* symbols, such as parts of speech, and their interactions with other terminal and nonterminal symbols are of interests for natural language processing. For example, production rules are used to characterize formal grammars. Stochastic trainable grammar was first proposed by Baker in 1979 [4]. The so-called *inside-outside* algorithm for context-free grammar (as opposed to the so called *forward-backward* algorithm for finite state grammar), has been used to train production rule probabilities and to perform stochastic parsing for both ASR and NLP tasks (e.g. [19, 40, 70]).

New stochastic modeling techniques have been proposed to handle word and sentence alignment of parallel text from two natural languages ([10]). Techniques have also been applied to characterize the undelying language structures and their relationship between two languages. For example, translation models between two natural languages (English and French) have been constructed entirely from the training corpora [10]. Grammatical association techniques [53, 76]. have also been proposed to construct stochastic transduction models from a stochastic grammar of a natural language to a stochastic grammar of another artificial language.

3.3 Adaptive Modeling of Linguistic Units

Most spoken language systems rely on a static design strategy in that all the knowledge sources needed in a system are acquired at the design phase and remain the same during testing. Since the samples used in the design are often limited, this results in some mismatch problems. A better way is to acquire the knowledge dynamically. New information is constantly collected during the testing stage and is incorporated into the system using adaptive learning algorithms.

For adaptive modeling of N-grams, some approaches have been proposed recently. The first uses a *cache* [34] obtained from performing the actual task. Usually a history of the last few hundred words is maintained and used to derive a cache trigram and then combine with the static trigram. This results in an adaptive trigram which is the weighted interpolation of both the static and the cache trigram. This technique is similar to the Bayesian adaptation technique ([20]) that combines the new observed data and the existing model in a maximum a posteriori sense.

The cache based approach can be extended to include long-distance dependency between words appearing in the training text. One such approach is the so-called *trigger-based modeling* ([41]) in which trigger word pairs are established in the training phase. Words appearing in the cache (the history) of the current task are used as triggers to modify the N-gram word probabilities of the words triggered by those trigger words. The *maximum entropy* principle is then used to update the N-gram word probabilities [41]. The MAP principle (e.g. [21]) and the *minimum discriminant* estimation [18] can also be used. To enhance spoken dialogue system design we expect more research be conducted in the area of adaptive modeling of underlying linguistic structures.

3.4 Emerging Modeling Techniques

With the availability of advanced data-driven approaches, such as the HMM and the ANN modeling frameworks, it is now relatively easy to design a speech recognition system as long as a large body of training data is available and a task specification is given. However, it is not possible to collect data that cover all the task variability. Therefore, task-specific database collection is not the ideal way to deal with the variability problems in stochastic modeling. Databases should be collected with the purpose of learning about the source of variability so that algorithms can be designed to identify and properly handle such variability.

Advances in language modeling are needed to improve the efficiency and effectiveness of large vocabulary spoken dialogue 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. Discriminative modeling approaches, such as minimum classification error training ([36, 35, 72, 73]) and error correcting grammatical inference (ECGI, [59]), are also useful in producing more accurate models even when the amount of language training data is only limited. Such a discriminative parameter learning strategy has been recently applied to create integrated speech and language models for Mandarin speech recognition [11]).

4 ASR and NLP Integration

As seen in Figures 1 and 2, many knowledge sources are needed for automatic speech recognition and natural language processing in order to find the most likely recognized sentence and the determine.the most appropriate action in a spoken language system. There are two basic heuristic search strategies, the *modular* and *integrated* approaches, to find the the most likely sentence that satisfies all the task constraints. We now describe issues related to the integration of speech and language knowledge sources in spoken language system design. We also discuss the performance tradeoffs among various search and decision strategies.

4.1 Integrated Search Strategy

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 a finite state representation of acoustics, lexical knowledge, syntax and semantics, it is possible to compile all the above knowledge sources into a single finite state network composed of acoustic states and grammar nodes and their connections. 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 probabilities cannot easily be 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.

4.2 Modular Search Strategy

On the other hand, for the modular approach, the recognized sentence is 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 to the above advantage, modular approaches are usually more computationally tractable than integrated approaches. However one of the major limitations with the modular approaches 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.

4.3 Multi-Pass Decision Strategies

As opposed to the traditional left-to-right, one-pass search strategies, multiple-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. The tree-trellis algorithm [71] and the forward-backward search [1] are two examples of such multi-pass strategies. These algorithms are usually designed to provide N-best string hypotheses (e.g. [65, 71]). To improve flexibility, simpler acoustic and language models are often used in the first pass as a rough match to produce a *word lattice*. Detailed models and detailed matches are applied in later passes to combine partial theories into the recognized sentence. Segment and phone lattices, followed by lexical access and language analysis, has also been successfully implemented (e.g. [46, 81]).

4.4 The N-Best Search Paradigm

The N-best search paradigm [66] is an ideal way for integrating multiple knowledge sources. It has been used for rescoring a preliminary set of candidate strings with higher-level constraints like a digit check-sum [71], with detailed cross-word unit models and long-term language models [5, 66], with segmental neural nets [79] and with prosodic models [75] for speech understanding, and with a semantic post-processor [55] for incorporating a full task grammar, etc. It has also been used to provide competing string hypotheses for discriminative training and for combining multiple acoustic models to reduce recognition errors [13]. We expect to see more use of the N-best search paradigm for designing large vocabulary speech recognition, speech understanding, spoken dialogue and speech translation systems.

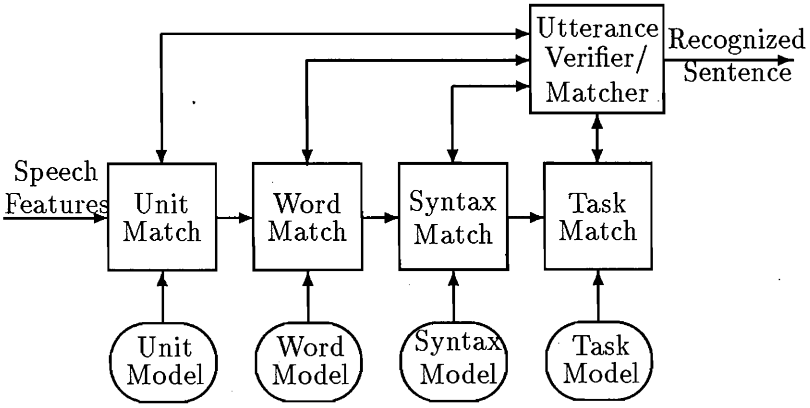

Figure 3: Block diagram of a top-down knowledge source integration.

4.5 Hybrid Search Strategy

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 spoken language tasks. Good *delayed decision* strategies in each decoding stage are required to minimize errors caused by hard decisions. Multiple word and string hypothesization is also crucial for the integration of multiple and sometimes incompatible knowledge sources. An example is shown in Figure 3 in which the partial theories from each matching module are integrated to find the recognized sentence through an utterance verifier.

5 Spoken Language Understanding

Speech recognition is the process of converting an input speech utterance into a sequence of words. To further capture the meaning of the recognized sentence requires the characterization of semantics behind the message. However meaning in natural language is often ambiguous. For database access tasks such as the DARPA ATIS task [27, 28] the required semantic knowledge is usually compiled manually. New rules are added based on the notion of reducing the errors in the training examples. Since different task-specific knowledge is needed for different tasks, the rules designed for one task usually cannot be generalized to another task. To circumvent the above difficulty, *task portability* research should be focused on a methodology that will generalize easily. Spoken dialogue systems have also been designed to study other research issues such as discourse analysis, dialogue modeling, natural language understanding and language generation.

In Table 1, we summarize some of the issues related to speech recognition and understanding. We use the DARPA Naval Resource Management task and the ATIS task as examples. It is clear that speech understanding system addresses some of the fundamental research issues of human-machine communication. First, the task vocabulary is often open that the user is not constrained to stay within a fixed vocabulary. The speech format is usually spontaneous with an implied spoken disfluency. Detection and rejection of extraneous speech events become essential in spoken language system design. New words unknown to the spoken system need to be detected and interpreted. Second, strict syntactic analysis is now error-prone because the text output from the recognizer is often ill-formed. Grammars for spoken language are also ill-defined and flexible grammar

Table 1: Summary of Speech Recognition and Understanding Issues.

Example	RM	ATIS
Goal	Recognition	Understanding
Focus	Acoustic Modeling	All Aspects
Vocab. Size	1000 Words	Open
Input	Read Speech	Spontaneous
New Word	None	Plenty
Syntax	Rigid	Flexible
Semantics	Not Used	Essential
Discourse	Not Used	Essential
Evaluation	Well-Defined	Open
Integration	Little	Plenty
Portability	Easy	Difficult
Expectation	5% Error	15% Error

should be used to accommodate for possible incomplete specification of syntactic rules. Third, semantic analysis becomes an essential system component. Research in knowledge representation, language acquisition and modeling of meanings is crucial to the success of spoken language systems. Fourth, discourse analysis and dialogue modeling play an important role in designing effective interface to enhance human-machine communication.

5.1 Domain-Specific Speech Understanding

Since text understanding of an unrestrict domain is not a solved problem, most of the speech understanding scenarios being considered in the research community are all in limited domain areas, including airline/train information access and flight/train/hotel reservation. The reader is referred to a pioneering example of the MIT VOYAGER system [81], which is a spoken language system for guiding users to find places in Cambridge. It has a multimedia output, including speech, graphics, map and text feedback to help the user. Multi-lingual VOYAGER systems have also been developed (e.g. [23]). The same strategy has also been extended to the MIT PEGASUS system, which is a spoken language interface, connecting to an on-line flight database to enable users to book real flights [82]; and the MIT GALAXY system, which is a telephone-based spoken language interface for accessing on-line information [24].

Conventional text based understanding approaches are predominantly syntax-driven (e.g. [67]). They usually assume that a complete set of grammatical rules can be compiled to perform whole-sentence parsing. Since the text input to the language analyzer in a spoken language system is typically ill-formed because it is produced by an imperfect speech recognition system, syntax-based approaches often fail miserably because a complete parse of the text is usually difficult, if not possible, for spoken inputs.

Because of the above limitations, some researchers have adopted semantics-driven approaches (e.g. [77]), relying on the detection of meaningful words and key phrases that bear semantic cues and ignoring events that are not relevant to understanding. Although such approaches may not be applicable to analyzing complex languages, they offer a good compromise in dealing with domain-specific, spontaneous utterances. Another alternative is to allow extraction of partial or fragmental parses and then glue them together to form a sentence level full or partial understanding (e.g. [68]). Both of the semantics-driven and *robust parsing* strategies make it east to achieve some form of rule-based text understanding via converting the text information into meaningful messages usually represented by a *semantic frame* (e.g. [31]). One can imagine the semantic frame being a structure that contains all the information needed to achieve a complete understanding. For example, in flight reservation, the semantic frame typically consists of all the information required to issue a airline ticket.

Recently, new stochastic semantic modeling algorithms have been proposed to learn semantic models and constraints from examples (e.g. [39, 22, 48, 55, 69]). Following the formulation in [55], we next illustrate how to design an integrated spoken language understanding systems.

5.2 Pattern Recognition Approach

We assume a source-channel speech generation model shown in Figure 4, in which the message source, M, contains meaning concept, C, and an associated sequence of words, W. Because of the uncertainty and inaccuracy in converting from message to speech, S, we model the conversion process as a noisy channel. Message understanding is then

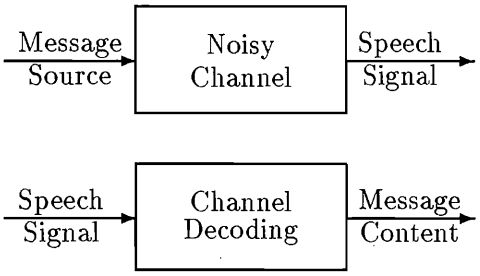

Figure 4: Source-channel model of speech generation and message understanding

formulated as a *maximum a posteriori* (MAP) decoding problem, as shown in Figure 4. 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,

$$\underset{M \in \Gamma}{\operatorname{argmax}} P(M|A) = \underset{M \in \Gamma}{\operatorname{argmax}} P(A|M) \cdot P(M), \tag{1}$$

where Γ is the set of all possible messages, $P(A|M)$ is the conditional probability of the acoustic vector sequence, A, given a particular message M, and $P(M)$ is the a priori probability of generating the particular message M. We assume M can be characterized as a pair of concept and word sequences, (C, W), where W is the sequence of words associated with the speech or acoustic signal, A, and C is the concept sequence embedded in the message with each concept c_i in C corresponds to the word w_i in W. Since each concept represents a semantic attribute for the particular domain of the message, C can be used to infer the meaning embedded behind the message with W encoding the value of the semantic attribute for further processing. For example, a particular c_i could represent the time concept and the word w_i="3pm" indicates a particular time, three o'clock in the afternoon, is referred to in the message. We can even encode the phrase "three o'clock in the afternoon" as "time.15". Therefore the problem becomes finding the most likely pair of concept and word sequence such that

$$\underset{M \in \Gamma}{\operatorname{argmax}} P(M|A) = \underset{M \in \Gamma}{\operatorname{argmax}} P(A|W) \cdot P(W|C) \cdot P(C) \tag{2}$$

where $M = (C, W)$. In the first term of the right-hand side of 2 we assume the acoustic signal only depend on the words being uttered and is independent of the concept being expressed. This gives the first term, $P(A|W)$, which is often referred to as an *acoustic* *model* in speech recognition. The second term, $P(W|C)$, is referred to as a concept-specific em language model. The last term. $P(C)$ is called a *concept model.*

The noisy channel in Figure 4 is a model jointly characterizing the message generation mechanism, the speech production system, the speaker variability, the speaking environment, and the transmission 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 $P_\theta(A|W)$, $P_\gamma(W|C)$ and $P_\omega(C)$, i.e. according to specific models. All the parameters of the statistical models (i.e. θ , γ and ω) needed in evaluting the acoustic probability, $P_\theta(W|A)$, the language probability, $P_\gamma(W|C)$, and the concept probability $P_\omega(C)$ are usually estimated from a large collection (the so-called *training set*) of speech, text and concept-annotated training data.

Detail of the above formulation and the application to the ATIS task can found in [55]. Extensions to it can also be found in [69, 48]. The readers are also referred to other pattern recognition approaches to speech and language processing problems, including speech recognition (e. g. [61]), part of speech tagging (e. g. [14]), machine translation (e.g. [10]), and integrated speech and language knowledge sources (e.g. [11]).

5.3 Robust and Flexible Speech Understanding

As spoken language systems or spoken dialogue systems are being evaluated for wider usage in real-world applications, it is found that they are not sufficient to cope with the utterance variation inherent in a large user population.

Conventional spoken language systems try to decode the whole input utterance with a pre-defined set of task constraints, including the fixed task vocabulary and the language models, and match every part of the input uniformly. For in-grammar sentence patterns, the use of a rigid task grammar, which is compiled from a set of task knowledge and real examples, is usually quite effective. However, in real-world environments, we have observed a large number of out-of-grammar utterances even after the task grammars had been tuned by human experts during the trial period. These samples include extraneous words, hesitations, repairs and unexpected expressions. There are also cases the user utterances are out-of-task, i.e. they have nothing to do with the task. Such utterances are usually difficult to identify and they make the systems hard to respond properly.

In the meantime, most of the mis-recognized utterances contain some key phrases that are task-related and may lead to partial or full understanding. Flexible speech understanding should be able to detect semantically significant parts and reject irrelevant parts. In a specific task domain of a transaction or information retrieval system, it is possible to make sense with key words or phrases. Therefore, an approach based on their detection is attractive. By relaxing the grammatical constraints and focusing on the key words and phrases, it will accept a wider variety of utterances than rigid sentence grammars can.

Such a detection-based speech understanding system consists of the following steps [38]:

1. *Key Phrase Detection:* A set of key phrases are detected using a set of phrase subgrammars specific to the dialogue state. The key phrases are labeled with semantic tags, which are useful in the sentence-level parsing and they lead to a direct sentence-level understanding.

2. *Key Phrase Verification:* The detected key phrases are verified and assigned *confidence measures.* This process eliminates false alarms and rescores the verified candidates. The verifier is constructed with acoustic subword and *anti-subword models* which test the individual subwords of the recognized results [74].

3. *Sentence Parsing and Detection:* The key phrase candidates are connected into sentence hypotheses using task-specific semantic knowledge sources. A parser [38] is used to construct string hypotheses that satisfy the semantic constraints.

4. *Sentence Verification and Rescoring:* The semantically valid sentence hypotheses are verified and rescored with detailed classification and verification models by reprocessing the speech input.

The above robust understanding strategy has been tested on utterances collected in a real-world application trial. Three categories of examples, in-grammar, out-of-grammar, and out-of-task utterances, have been evaluated. When compared with results obtained with the conventional approach of using a set of rigid grammars, we found the detection-based understanding approach gives about the same semantic frame recognition accuracy for in-grammars utterances. Furthermore, it achieves a much better semantic frame accuracy for both out-of-grammar and out-of-task utterances. However, more research is required to improve further the handling of ill-formed utterances [38].

6 Real-World Spoken Dialogue: Research Issues

We now address issues related to designing real-world spoken dialogue systems. We use a voice-based car reservations (VBCR) system as an example. The VBCR system is a prototype trial system for making car reservations by phone via speech input. Speech output is also used to generate voice prompt to communicate with the user. The information provided by the user includes account number, pick-up and drop-off locations, date, time, and flight numbers. It can be considered as a *voice form filling* application that allows users to fill all the fields on a form with speech utterances containing field-specific information. Two parallel efforts were carried out. The first was to implement such a prototype using a commercially available ASR software package [47]. The other is to start with a research algorithm and conduct a study on what's needed to realize such a real-world application.

The original intent was to let the user fill in, at a time, as many fields in the reservation form as possible based on a *mixed-initiative* dialogue strategy similar to what's adopted in the ATIS evaluation. Since there were not enough application-specific dialogue examples to train the language models for speech recognition and for dialogue processing, it was soon discovered that mixed-initiative dialogue gave a poor performance with the commercial ASR package. Three key improvements were soon adopted. First, a large set of telephone based training data was provided to the software provider to train an improved set of subword models. Second, the dialogue was constrained so that the user only fill in one field at a time. The constraints were imposed through a set of field-specific voice prompts. Third, the grammar for each field was manually adjusted based on the trail sample collected and it was defined using a deterministic finite-state grammar. The reason for this is partly due to the realistic limitation that it is not possible to have collected enough application-specific dialogue examples to build reliable stochastic language models and dialogue models.

In contract to the mixed-initiative strategy, this is a *system-initiative* dialogue strategy, similar to what's adopted in the MIT PEGASUS [82] system, which is more appropriate for this type of real-world applications. In a system-initiative dialogue mode, the system expects the user to provide specific information. The dialogue between the user and the system is therefore somewhat restrictive in the sense that the system initiates the dialogue by soliciting this information from the user through a field-specific voice prompt request. Usually, the prompt constrains the user voice input so that the dialogue is carried out in an unambiguously manner. For experienced users, this is an efficient way to get the job done in a short amount of time (e.g. 2 minutes per transaction session). However, this type of dialogue is usually not natural. It sometimes leads to unrecoverable difficulty and frustration for inexperienced users.

Based on our parallel study in enhancing our research algorithm to handle such a system-initiative dialogue strategy, we found the following: (1) It is not easy to have available a large set of dialogue examples to design reliable stochastic language models for every new dialogue application; (2) Semantic tagging for each new application is labor-intensive; (3) Domain coverage based on a small set of dialogue examples is poor; (4) High performance task-independent acoustic models are required for robust recognition of utterance in different tasks; (5) Task-dependent acoustic models, such as alphabet and digit models, give better performance than task-independent models when the test data are from these tasks (e.g. alpha-digit recognition); (6) High performance confidence measures are required to verify if the input utterances are valid so that an intelligence speech interface can be designed to help with voice repairs and confirmation and rejection of invalid input; (7) Human factors research is essential for enhancing spoken language system usability and portability. Although some of the abovementioned issues can be addressed with the robust speech understanding algorithm discussed in the previous Section, there are still many open research issues. We briefly state some of them in the following. Hope this will inspire new research directions.

6.1 Continuing Research Issues

It is clear that spoken language system research encompasses automatic speech recognition, natural language processing and human-machine interface technology. Many researchers have issues challenges to this exciting new research area. A good example can be found in the collective report by Cole *et al* [15].

In order to realize a usable spoken language system, a number of new research advances are needed. These enhancements can be summaried in four key areas, namely: (1) improving ASR performance; (2) handling flexible subgrammars; (3) incorporating utterance verification; and (4) automatic task and dialogue generation with improved semantics and language processing. We discuss these enhanced features in more detail in the following.

6.1.1 Improving ASR Performance

- *Task-dependent and task-independent acoustic modeling*: It is well known that task-dependent training usually achieves the best performance for a specific task. However, it is not possible to collect a large amount of training data for every new task. For some common tasks, such as digit and alphabet recognition, it is always beneficial to have models trained on existing large databases. This can also be extended to other useful tasks such as date, time and natural number recognition. Language models can also be trained this way. However, general acoustic models that can be used for other tasks with using specific task knowledge are also important. Research here includes [44]: (1) the design of general training databases for all general tasks; (2) fundamental unit selection and modeling strategy that can be generalized to many tasks; (3) how to learn from the vocabulary and grammar of a particular task and design the acoustic models accordingly without the need to collect examples of speech data for the task; and (4) how to adapt to a particular task based on a small set of acoustic and language models.

- *Robustness improvement*: Recognition performance of a system is often degraded due to mismatch conditions existing between training and testing. Robust feature extraction, signal conditioning, speaker normalization, and speaking environment compensation are a few things that can be incorporated to improve system performance. Real time implementation issues need to be considered also when robust techniques are being incorporated.

- *On-line speaker adaptation*: Since a user is likely to stay on a dialogue system longer, some of the recognized utterances in the earlier part of the dialogue session can be used to perform speaker and environment adaptation so that the recognition performance can be improved for the later part of the dialogue session. On-line unsupervised adaptation needs to be done here because it is not easy to acquire supervision information in an operational system environment.

6.1.2 Handling Flexible Subgrammars

- *Definition of localized subgrammars*: Instead of using the current FSG's, the embedded constraints in FSG should be relaxed so that ungrammatical utterances can also be handled. Each localized subgrammar is typically composed of keywords, key-phrases and their contexts. They are semantically tagged and designed to handle only the relevant semantic events in a particular field, e.g. date, time and account number recognition. Stochastic FSG's can also be used if we have enough training data to train them.

- *Keyword and phrase detection based on localized subgrammars*: Detecting keywords and key-phrases in a subgrammar is usually easier than straight recognition. High detection accuracy can be achieved even with less accurate acoustic and language models. However, reducing false alarms is an importance research topic. This can be done with the combination of utterance verification and task-specific semantic constraints.

- *String hypothesization and verification based on connected subgrammars*: The above semantically tagged subgrammars can be connected to form meaning string hypotheses. They are then verified using the whole speech utterance to choose the most likely string that is also semantically meaningful. Many subgrammars can also be connected into a single recognition grammar using a set of relaxed global semantic constraints so that multiple fields can be filled in with a single spoken utterance.

- *Automatic subgrammar and vocabulary generation*: The current design procedure starts with a given task definition and constructs all subgrammars and vocabulary words manually based on human expert knowledge. This is a time-consuming process especially for defining new tasks. It is also highly error-prone. Research is needed in the area of automatic subgrammar and vocabulary generation. Task representation (could be from database schema) and interface between task representation and generation needs to be established so that task mapping to spoken language systems can be simplified.

6.1.3 Incorporating Utterance Verification

- *Utterance verification to produce word and phrase level confidence*: Instead of doing the conventional one-pass recognition, which was designed to handle well-formed utterances, recognition for ill-formed utterances should be done by keyword and phrase detection followed by utterance verification. Issues related to task dependency and verification strategies need to be studied.

- *Robust hypothesis pruning based on partial utterance verification*: As discussed above. keyword and phrase detection is likely to produce a large number of false alarms. Utterance verification is a robust way to perform hypothesis pruning. The word and phrase level confidence scores are useful for reducing false alarms.

- *Voice repair and prompt generation based on phrase confidence*: The word and phrase level confidence scores are key indicators about how well an utterance is being recognized. An intelligent user interface can use this information to decide the level of voice confirmation, how much to prompt, what to prompt, and when the user needs help through voice repair.

6.1.4 Semantics and Language Processing

- *Semantics dictionary*: In order to accomplish some form of understanding, more research in semantics representation and modeling is needed. Some research tools, such as a large machine-readable dictionary with semantic definitions for words, are keys to the advancement in this key area.

- *Language and knowledge acquisition*: Human beings acquire language and related knowledge sources in a natural way through exposure to the vast learning environments offered in their families, schools and societies. It is important for machines to be quipped with learning capabilities to acquire language either automatically or semi-automatically. This set of knowledge is crucial for designing large-scale spoken language systems.

- *Portability research*: Up to this point most of the dialogue systems are designed manually based on human expert knowledge. Some of the rules generated in one task by a designer may not apply to other tasks or other users. It is desirable to have an ability for automatic dialogue generation based on task definition. This is an important portability research topic.

7 Summary

We have briefly reviewed the present state of spoken language processing. An extensive list of references are also included for further reading although some of them are not referred to in the text of this paper. In recent years, we have learned a great deal about how to build and efficiently implement laboratory spoken language systems. However there remain a whole range of fundamental questions for which we have no definitive answers. Spoken language processing is an exciting research area. It is also a multi-discipline area that it is not easy for a single research group to solve all the problems. Collaborative research has shown to be a powerful way to enhance our current capabilities. In order to make progress in Chinese speech and language processing, it is important for the research community to establish an infrastructure to develop large scale speech and language corpora, to share knowledge and tool resources, to improve communication among different groups, and to come up a good common problem that is challenging enough to warrant new research advances. It is also crucial to work on real-world problems because of the inherent nature of spoken language processing. This new area is still in its infancy. Many advances are yet to be made. It is up to this community to make contributions to the Chinese language processing part so that we can get connected to the brave new world of multi-lingual, human-human and human-machine communication.

References

[1] S. Austin, R. Schwartz and P. Placeway, "The Forward-Backward Search Algorithm," *Proc. ICASSP-91*, pp. 697-700, Toronto, 1991.

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

[3] L. R. Bahl, P. F. Brown, P. V. de Souza and R. L. Mercer, "Tree-Based Language Model for Natural Language Speech Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 37, No. 7, pp. 1001-1008, 1989.

[4] J. Baker, "Trainable Grammar for Speech Recognition," in *Speech Communication Papers for the 97th Meeting of the Acoustical Society of America*, J. J. Wolf and D. .H. Klatt, Editors, Cambridge, 1979.

[5] L. Bates, et al., "The BBN/HARC Spoken Language System," *Proc. ICASSP-93*, Vol. 2, pp. 45-48, Minneapolis, 1993.

[6] L. E. Baum, 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, pp. 164-171, 1970.

[7] J. R. Bellegarda and D. Nahamoo, "Tied Mixture Continuous Parameter Modeling for Speech Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 38, No. 12, pp. 2033-2045, 1990.

[8] R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, 1957.

[9] R. Bobrow, R. Ingria and R. Stallard, "Syntactic and Semantic Knowledge in the DELPHI Unification Grammar," *Proc. DARPA Speech and Natural Language Workshop*, Hidden Valley, 1990.

[10] P. F. Brown, et al., "A Statistical Approach to Machine Translation," *Computational Linguistics*, Vol. 16, No. 2, pp. 79-85, 1990.

[11] T.-H. Chiang, 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, pp. 167-189, 1996.

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

[13] W. Chou, C.-H. Lee and B.-H. Juang, "Minimum Error Rate Training Based on the N-Best String Models," *Proc. ICASSP-93*, Vol. 2, pp. 652-655, Minneapolis, 1993.

[14] K. W. Church, "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text," *Proc. Second Conference on Applied Natural Language Processing*, Austin, 1988.

[15] R. Cole, et al, "The Challenge of Spoken Language Systems: Research Directions for the Nineties," *IEEE Trans. Speech and Audio Processing*, Vol. 3, No. 1, pp. 1-21, 1995.

[16] H. D'Hooge, "The Communicating PC," *IEEE Communications Magazine*, pp. 36-47, Vol. 34, No. 4, 1996.

[17] S. Doshita, "Research on Understanding and Generating Dialogue by Integrated Processing of Speech, Language and Concept," *Research Report, 1993 - 1996*, Department of Information Science, Kyoto University, March 1996.

[18] S. Della Pietra, V. Della Pietra, R. L. Mercer and S. Roukos, "Adaptive Language Modeling Using Minimum Discriminant Estimation," *Proc. ICASSP-92*, pp. 633-636, San Francisco, 1992.

[19] T. Fujisaki, et al., "A Probabilistic Parsing Method for Sentence Disambiguation," *Proc. International Workshop on Parsing Technologies*, Pittsburgh, 1989.

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

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

[22] E. P. Giachin, "Automatic Training of Stochastic Finite-State Language Models for Speech Understanding," *Proc. ICASSP-92*, pp. 173-176, San Francisco, 1992.

[23] J. Glass, et al, "A Bilingual VOYAGER System," *Proc. EuroSpeech-93*, pp. 2063-2066, Berlin, Sept. 1993.

[24] D. Goddeau, et al, "GALAXY: A Human Language Interface to On-Line Travel Information," *Proc. ICSLP-94*, pp. 707-710, Yokohama, Sept. 1994.

[25] I. J. Good, "The Population Frequencies of Species and the Estimation of Population Parameters," *Biometrika*, Vol. 40, pp. 237-264, 1953.

[26] B. Grosz and C. Sidner, "Plans for Discourse," in *Intentions in Communication*, MIT Press, Cambridge, MA, 1990.

[27] C. T. Hemphill, J. J. Godfrey and G. D. Doddington, "The ATIS Spoken Language System Pilot Corpus", *Proc. DARPA Speech and Natural Language Workshop*, Hidden Valley, 1990.

[28] L. Hirshmann, et al, "Multi-Site Data Collection for a Spoken Language Corpus," *Proc. ICSLP-92*, pp. 903-906, Banff, Oct. 1992.

[29] J. Hobbs, "FASTUS: A Cascades Finite-State Transducer for Extracting Information from Natural-Language Text," *Proc. ROCLING IX*, Tainan, Taiwan, August 1996.

[30] X. Huang and M. A. Jack, "Semi-continuous hidden Markov models for speech signal," *Computer, Speech and Language*, Vol. 3, No. 3, pp. 239-251, 1989.

[31] E. Jackson, D. Applet, J. Bear, R. Moore, and A. Podnozny "A Template Matcher for Robust NL Interpretation," *Proc. DARPA Speech and Natural Language Workshop*. pp. 190-194, Pacific Grove, 1991.

[32] F. Jelinek and R. L. Mercer, "Interpolated Estimation of Markov Source Parameters from Sparse Data," in *Pattern Recognition in Practice*, E. S. Gelsema and L. N. Kanal, Editors, North-Holland Publishing Co., Amsterdam, 1980.

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

[34] F. Jelinek, B. Merialdo, S. Roukos and M. Strauss, "A Dynamic Language Model For Speech Recognition," *Proc. DARPA Speech and Natural Language Workshop*. pp. 293-295, Pacific Grove, 1991.

[35] B.-H. Juang and S. Katagiri, "Discriminative Learning for Minimum Error Classification," *IEEE Trans. Signal Processing*, Vol. 40, No. 12, pp. 3043-3055, 1992.

[36] S. Katagiri, C.-H. Lee and B.-H. Juang, "New Discriminative Training Algorithms Based on the Generalized Probabilistic Descent Method," *Proc. IEEE-SP Workshop on Neural Networks for Signal Processing*, pp. 299-308, Princeton, 1991.

[37] S. M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 35, No. 3, pp. 400-401, 1987.

[38] T. Kawahara, C.-H. Lee and B.-H. Juang, "Key-Phrase Detection and Verification for Flexible Speech Understanding", *Proc. ICSLP-96*, Philadelphia, Oct. 1996.

[39] R. Kuhn and R. DeMori, "Learning Speech Semantics with Keyword Classification Trees," *Proc. ICASSP-93*, Vol. 2, pp. 55-58, Minneapolis, 1993.

[40] J. Kupiec, "Hidden Markov Estimation for Unrestricted Stochastic Context-Free Grammars," *Proc. ICASSP-92*, pp. 177-180, San Francisico, 1992.

[41] R. Lau, R. Rosenfield and S. Roukos, "Trigger-Based Language Models: A Maximum Entropy Approach," *Proc. ICASSP-93*, Vol. 2, pp. 45-48, Minneapolis, 1993.

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

[43] C.-H. Lee and B.-H. Juang, "A Survey on Automatic Speech Recognition with An Illustrative Example on Continuous Speech Recognition of Mandarin," *Journal of Computational Linguistics and Chinese Language Processing*, Vol. 1, No. 1, Sept. 1996.

[44] C.-H. Lee, 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.

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

[46] S. E. Levinson, M. Y. Liberman, A. Ljolje and L. G. Miller, "Speaker-Independent Phonetic Transcription of Fluent Speech for Large Vocabulary Speech Recognition," *Proc. ICASSP-89*, pp. 441-444, Glasgow, 1989.

[47] S. M. Marcus, et al, "AutoRes System - Prompt Constrained Natural Language in Telephony Services," *Proc. ICSLP-96*, Philadelphia, Oct. 1996.

[48] S. Miller, R. Schwartz, R. Bobrow and R. Ingria, "Statistical Language Processing Using Hidden Understanding Models," *Proc. ARPA Speech and Natural Language Workshop*, March 1994.

[49] Y. Minami, K. Shikano, T. Yamada, and T. Matsuoka, "Very Large Vocabulary Continuous Speech Recognition Algorithm for Telephone Directory Assistance," *Proc. EuroSpeech-93*, Berlin, 1993.

[50] A. Nagai, S. Sagayama, and K. Kita, "Phoneme-context-Dependent LR parsing algorithms for HMM-based continuous speech recognition," *Proc. EuroSpeech-91*, pp. 1297-1400, Genova, 1991.

[51] H. Ney, "Dynamic Programming Parsing for Context-Free Grammar in Continuous Speech Recognition," *IEEE Trans. Signal Processing*, Vol. 39, No. 2, pp. 336-340, 1991.

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

[53] J. Oncina, P. Garcia and E. Vidal, "Learning Subsequential Transducers for Pattern Interpretation Tasks," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 15, No. 5, 1993.

[54] J. Peckham, "A New Generation of Spoken Dialogue Systems: Results and Lessons from the SUNDIAL Project," *Proc. EuroSpeech-93*, pp. 33-40, Berlin, Sept. 1993.

[55] R. Pieraccini, E. Levin and C.-H. Lee, "Stochastic Representation of Conceptual Structure in the ATIS Task," *Proc. DARPA Speech and Natural Language Workshop*, pp. 121-124, Pacific Grove, 1991.

[56] R. Pieraccini and C.-H. Lee, "Factorization of Language Constraints in Speech Recognition", *Proc. ACL-91*, Berkeley, 1991.

[57] P. Placeway, R. Schwartz, P. Fung and L. Nguyen, "The Estimation of Powerful Language Models from Small and Large Corpora," *Proc. ICASSP-93*, Vol. 2, pp. 33-36, Minneapolis, 1993.

[58] P. Price, "Evaluation of Spoken Language Systems: The ATIS Domain," *Proc. DARPA Speech and Natural Language Workshop*, pp. 91-95, Hidden Valley, June 1990.

[59] N. Prieto and E. Vidal, "Learning Language Models through the ECGI Method," *Proc. EuroSpeech-91*, Geneva, 1991.

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

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

[62] D. B. Roe, et al., "A spoken language translator for restricted-domain context-free language," *Speech Communication*, Vol. 11, Nos. 2-3, pp. 311-319, 1992.

[63] A. Rudnicky, J.-M. Lunati and A. Franz, "Spoken Language Recognition in An Office Management Domain," *Proc. ICASSP-91*, pp. 829-832, Toronto, 1991.

[64] S. Sagayama, et al, "ATREUS: Continuous Speech Recognition Systems at ATR Interpreting Telephony Research Laboratories," *Proc. SST-92*, pp. 324-329, Brisbane. 1992.

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

[66] R. Schwartz, et al., "New Uses for the N-Best Sentence Hypotheses Within The BBN BYBLOS Continuous Speech Recognition System," *Proc. ICASSP-92*, pp. 1-4, San Francisco, 1992.

[67] S. Seneff, "TINA: A Natural Language System for Spoken Language Applications," *Computational Linguistics*, Vol. 18, No.1, pp. 61-68, March 1992.

[68] S. Seneff, "Robust Parsing for Spoken Language Systems," *Proc. ICASSP-92*, pp. 189-192, San Francisco, 1992.

[69] S. Seneff, H, Meng, and V. Zue, "Language Modelling for Recognition and Understanding Using Layered Bigram," *Proc. ICSLP-92*, pp. 317-320, Banff, Oct. 1992.

[70] R. A. Sharman, F. Jelinek and R. L. Mercer, "Generating a Grammar for Statistical Training,' *Proc. DARPA Speech and Natural Language Workshop*, Hidden Valley, 1990.

[71] F. K. Soong and E.-F. Huang, "A Tree-Trellis Based Fast Search for Finding the N-Best Sentence Hypotheses in Continuous Speech Recognition," *Proc. ICASSP-91*, pp. 705-708, Toronto, 1991.

[72] K.-Y. Su, T.-H. Chiang and Y.-C. Lin, "A Unified Framework to Incorporate Speech and Language Information in Spoken Language Processing," *Proc. ICASSP-92*, pp. 185-188, San Francisco, 1992.

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

[74] R. Sukkar 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.

[75] N. M. Veilleux and M. Ostendorf, "Probabilistic Parse Scoring with Prosodic Information," *Proc. ICASSP-93*, Vol. 2, pp. 51-54, Minneapolis, 1993.

[76] E. Vidal, R. Pieraccini and E. Levin, "Learning Associations Between Grammars: A New Approach to Natural Language Understanding," *Proc. EuroSpeech-93*, Berlin, Sept. 1993.

[77] W. Ward, "The CMU Air Travel Information Service: Understanding Spontaneous Speech," *Proc. DARPA Speech and Natural Language Workshop*, pp. 127-129, Hidden Valley, June 1990.

[78] S. Young and P.C. Woodland, "The Use of State Tying in Continuous Speech Recognition," *Proc. EuroSpeech-93*, pp. 2203-2207, Berlin, Sept. 1993.

[79] G. Zavaliagkos, 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, No. 1, pp. 151-160, Jan. 1994.

[80] V. Zue, S. Seneff and J. Glass, "Speech Database Development at MIT: TIMIT and Beyond," *Speech Communication*, Vol. 9, No. 4, pp. 351-356, August 1990.

[81] V. Zue, et al., "Integration of Speech Recognition and Natural Language Processing in the MIT VOYAGER System," *Proc. ICASSP-91*, pp. 713-716, Toronto, 1991.

[82] V. Zue, et al., "PEGASUS: A Spoken Language Interface for On-Line Air Travel Planning," *Proc. ARPA Speech and Natural Language Workshop*, March 1994.

[83] V. Zue, "Human Computer Interactions Using Language Based Technologies," *Proc. International Symposium on Speech, Image Processing and Neural Networks*, pp.i-vii, Hong Kong, April 1994.