Learning New Words from Spontaneous Speech: A Project Summary

Sheryl R. Young

July 29, 1993
CMU-CS-93-223

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

This research develops methods that enable spoken language systems to detect and correct their own errors, automatically extending themselves to incorporate new words. The occurrence of unknown or out-of-vocabulary words is one of the major problems frustrating the use of automatic speech understanding systems in real world tasks. Novel words cause recognition errors and often result in recognition and understanding failures. Yet, they are common. Real system users speak in a spontaneous and relatively unconstrained fashion. They do not what words the system can recognize and thereby are likely to exceed the system's coverage. Even if speakers constrained their speech, there would still be a need for self-extending systems as certain tasks inherently require dynamic vocabulary expansion (e.g. new company names, new flight destinations, etc.). Further, it is costly and labor intensive to collect enough training data to develop a representative vocabulary (lexicon) and language model for a spoken interface application. Unlike transcription tasks where it is often possible to find large amounts of on-line data from which a lexicon and language model can be developed, for many tasks this is not feasible. Developers of applications and database interfaces will probably not have the resources to gather a large corpus of examples to train a system to their specific task. Yet, most current speech and language model research is oriented toward training from large corpora. This research enables systems to be developed from small amounts of data and then "bootstrapped". A simple version of a system is initially created using only a small set of examples. Its coverage is then automatically extended as the system encounters new input through interaction with users.

The system learns new words from spontaneous speech and incrementally augments its grammar and lexicon. Specifically, it detects new words and phrases, spells the new words, determines their meaning and adds them to the system lexicon and grammars. Lexical representations are initially created to permit recognition of the new words and phrases in the original dialog. Later representations are generalized to ensure robust recognition by different speakers in different syntactic and semantic contexts. There are two important aspects of our approach:

- **Conjoined use of many knowledge sources to detect recognition errors and determine whether an out-of-vocabulary word is present.** The interaction of these knowledge sources is critical in the success of the effort.
- **The system uses its higher-level knowledge sources to determine the meaning of new words to determine how to add the new words to the system grammar, language models and lexicon.** The system is able to detect new words that are similar to known words when they have different meanings. Similarly, the system is constrained from overgenerating new words as similar words that have the same meaning are merged in the robust lexical generalization phrase when they have be permitted to be hypothesized in the first place.

The system uses a lexicon, a dictionary, stochastic grammars, semantic grammars, acoustic models, a semantic domain knowledge base, a domain plan tree, pragmatics and dialog structure. In particular, this research is novel in that it emphasizes pragmatics, dialog structure and discourse level constraints. The research is being developed and demonstrated using speaker-independent, continuous, spontaneous speech.

1 Learning from Mistakes and New Word Acquisition: Objectives

It has long been the goal of speech and artificial intelligence researchers to build truly adaptive, robust and intelligent computer interfaces that permit users to interact with computer systems / programs in a natural and unconstrained manner, for example using spoken discourse. One version of such a system is an intelligent spoken interface to one or more computer applications. Here users spontaneously and naturally converse with a computer that assists the user in accomplishing their objectives by retrieving and interpretting stored information, initiating computer processes and executing commands, as appropriate. Such a system must be automatically extensible and incrementally adapt to the user and environment. It needs to determine when it has not correctly recognized or understood the input and learn from its mistakes, incrementing its lexicon (the words it knows) and domain knowledge as required.

There are two reasons why the ability to detect recognition errors and automatically correct them is essential for spoken language systems to become a viable, truly usable technology:

1. It is not possible to anticipate all the words that may be spoken and the system "breaks" when new words are encountered.
2. It is impractical and expensive to collect the requisite amounts of training data needed today for each potential application domain.

First, it is not possible to anticipate all the words a user may say even when very large amounts of training data are collected for each spoken language application. New words are frequently present in spontaneous speech. Real users do not know a system's lexical limitations. Inexperienced and casual users have no way to know what words a system recognizes and experienced users will not remember every lexical constraint. Futher, some tasks inherently require dynamic vocabulary expansion, for example, a travel planning system will need to add locations while a corporate tracking system will need to add new company names.

The occurence of unknown words is one of the major problems frustrating the use of speech understanding systems in real world applications because today, speech recognition technology is based upon matching input sound representations with only the words in a known, finite, and often quite limited lexicon. This is an underlying constraint on the techniques currently employed. Speech understanding minimally requires recognition of word meanings. In a spoken language application system, not only must words be correctly recognized, they must be understood and used to interpret the meaning of the utterance, mapping it into one or more system actions. The presence of novel words results in a series of word substitution errors and frequently causes a complete failure of recognition and understanding. Hence, for speech to become a truly usable medium for real application systems, it will be necessary to be able to detect novel word strings, determine the meaning of the words and add them to the system automatically.

Second, spoken language application systems cannot be trained comparably to speech recognition or word spotting applications. The training data are constrained to the application domain and are thereby limited. Further, the collection of training data is generally very expensive and these costs are significantly greater in spoken language understanding applications, or those areas were speech technology is most likely to have the most significant impact in the near future. In contrast to recognition systems that can be trained by having speakers read sentences, these systems must be trained by pre-defining and simulating anticipated system functionality and asking potential users to solve some set of problems in the application domain. The data obtained are limited and less likely to be comprehensive, as they reflect any biases in data collection procedures, including the form and content of information displayed, temporary limitations on the application (usually the scope of the application is incrementally defined or broken down into "manageable parts" during the development cycle), and the methods of responding to out-of-domain or ambiguous requests. For example, the presence or absence of abbreviations in database responses significantly altered the types of spoken behavior observed in ARPA's ATIS training data. The training data are very sensitive to minor modifications in the data collection procedures.

The dependence of today's spoken language systems upon training data is profound. Training data are used to define and train the system lexicon and the language model, requisites to accurate recognition. They are used to define and develop the system grammars and semantics, necessary to infer utterance meaning. To make spoken language application systems viable, robust and affordable, we need to develop methods for "bootstraping" applications from limited amounts of training data, correcting and expanding the domain knowledge sources in response to previously unencountered input.

The development of a system that can automatically learn and increment its coverage in response to interaction can be broken down into six ordered stages. Each stage builds on prior stages, and brings significant increases in functionality, robustness and adaptibility while significantly reducing the constraints placed upon the speaker.

- **Recognition:** The first, requisite stage is to achieve reasonable recognition rates on moderately scoped application domains. Reasonably accurate and reliable recognition is a prerequisite to a useful, usable spoken language system.
- **Meaning Inference:** The second stage is the ability to infer the meaning of a spoken utterance, taking into account the information entailed from contextually applicable preceeding discourse. Prior discourse knowledge can also be used to constrain the recognition and interpretation processes, thereby enhancing accuracy.
- **Spontaneous Interaction:** The third stage is to permit natural spontaneous interaction, inclusive of spontaneous speech and dialog. Spotaneous speech is inherently noisy containing acoustic, syntactic and semantic irregularities (relative to read or written input) that challenge systems by significantly increasing the scope of phenomena they must process. The search space for words must include "verbal noise", syntax is regularly violated and semantic ill-formedness is common. Unconstrained dialog includes clarification, confirmation and correction subdialogs as well as spontaneous topic shifts. Permitting the user to speak naturally and spontaneously removes constraints on the speech itself as well as on the structuring and ordering of words and utterances.
- **New Word Acquisition:** The fourth stage removes the constraints of a fixed, predefined lexicon, enabling novel words and phrases to be incrementally recognized and added to the system's lexicon, language model, grammar and domain semantics. This stage removes the requirement that speakers use a fixed word set in a predefined or grammatically anticipated manner, significantly increasing the flexibility and usability of spoken interface applications.
- **Learning from Mistakes:** Beyond robustness and real world usability, an adaptible interface not only acquires novel word strings but is capable of more generally self-diagnosing and correcting its errors. The stage spans a wide variety of behaviors beginning with detecting recognition errors, inadequate lexical representations and insufficiencies in a statistical language model or inadequate grammatical coverage. More complex learning includes detecting and correcting interpretation errors, database inconsistencies, inadequate database retrieval or database interface procedures as well as the ability to acquire new plan and action scripts. Finally, an adaptive system will learn user preferences such as preferred methods of responding to specific input queries, unanswerable requests or user plan failures.
- **New Concept Acquisition:** The sixth stage in developing a truly robust, incrementally adaptive spoken language system permits a system to incrementally acquire new concepts. This stage enables the scope of a spoken language application domain to be automatically expanded through interaction with a user.

The inception of the ARPA Spoken Language Understanding project has resulted in the investigation of stages one through three. Baseline recognition accuracy is steadily improving. Emphasis has shifted from recognizing read speech to understanding spontaneous speech and responding to spoken instructions. The technical challenges posed by spontaneous speech are being steadily conquered. Recognizers process extraneous acoustic information such as filled pauses, partial words and stutters [Ward, 1989, Wilpon, 1990]. Robust parsers are being developed to process dysfluent input, mid-utterance corrections and restarts. Spoken language systems that map spoken input into system actions are wide spread in the ARPA community [CMU, ATT, BBN, SRI]. More over, we have developed methods that permit spoken language systems to capitalize upon higher-level knowledge sources to enahnce recognition and interpretation accuracy [Young, 1989, Young 1990, Young and Ward, 1993]. These systems use dialog structure, semantics and pragmatics to compute constraints on what can logically be said next. This information has been used to dynamically modify langauge models and grammars [MINDS] and to detect and selectively reprocess errorful substrings of input [MINDS-II]. Basically, the recognition and understanding processes are constrained to consider those hypotheses that make sense.

We intend to build an experimental system that will enable us to develop and compare new techniques that permit spoken language understanding systems to recognize and understand new words from spontaneous spoken dialog interactions. Moreover, the techniques that enable automatic word learning are more general, permitting systems to begin to detect when they have made a mistake and to automatically extend their knowledge sources to prevent future, similar errors. We will develop methods that enable systems to detect recognition errors, methods for determining the likely source of the error and methods for automatically extending a system's lexicon, language model and grammar. The research emphasizes the conjoined use of acoustic, semantic, pragmatic, discourse structure and dialog-level knowledge and constraints to detect errors and infer the meaning of novel words. The proposed system will capitalize upon our existing in-house speech recognition and understanding systems and our prior work on spontaneous speech and dialog. We can benefit greatly from the infrastructure (recognizers, speech understanding systems, databases, etc.) and research resulting from our ARPA-funded work. Building upon existing capabilities greatly reduces the overhead of the proposed endeavor, permitting us to directly focus on the underlying scientific and technological issues and thereby speeding system prototyping, algorithm exploration and experimentation.

2 The Missing Science

The proposed research explores new techniques that permit a system to determine that it has made an error and to automatically extend itself to correctly process the input in the future. The system will detect recognition errors, determine their cause --distinguishing novel words from other error sources--, and automatically modify or increment its knowledge sources. When encountered, new words and phrases will be represented acoustically and added to the system lexicon, language model and grammar once their meaning has been inferred. Errors caused by other knowledge source failures, such as inadequate grammatical coverage, new word senses, misrepresentative language models and poor acoustic models will be detected and corrected incrementally.

While this research is driven by functionality considerations, taking a major step beyond all currently funded ARPA research, and providing for scientific evaluation of the research results, our approach entails addressing the following scientific gaps in current technology. These four additional contributions can be applied to more generally enhance the current state-of-the-art in speech recognition and speech understanding.

- **Confidence measures to assess recognition accuracy:** To develop a model for speech recognition that is able to detect and delimit unknown words minimally requires that we be able to detect misrecognized words. Today's recognizers evaluate competing word hypotheses against one another, taking into account statistical language model probabilities. They do not independently measure goodness of acoustic match. We will investigate alternate methods for evaluating acoustic goodness of match by developing acoustic normalization techniques and experimenting with alternate techniques for estimating independent acoustic confidence measures. Further, we will investigate methods for using higher-level knowledge sources with the acoustic confidence measures to develop a more reliable measure of recognition accuracy. These goodness of match assessments are independent of baseline recognition rates.
- **Techniques to estimate differential reliability and discriminative power of knowledge**

sources: To determine whether a word is misrecognized and to determine why, and whether a new word has been encountered, we want to use all of the knowledge available to a spoken language system, including dialog, semantics, pragmatics and acoustics. To know how to use these knowledge sources together, it is important to determine where they are most reliable and under what conditions they can reliably differentiate errors. For example, we know that when semantic, pragmatic and dialog-based knowledge flag an error, it is rarely a false alarm. However, these knowledge sources cannot detect errors where the word substitutions are contextually appropriate, 35% of the errors [Young, 1993d]. We will develop methods for assessing and storing meta-knowledge about the relative strengths and weaknesses of each knowledge source. These will record the number of times an error of a specific type is observed as well as the context under which it was observed.

- **Methods for merging knowledge sources that take into account differential reliability and discriminative power:** This question is essentially how basically different knowledge sources can be optimally combined. Bayesian updating is a well-known method for merging probabilistic information. However, some knowledge is optimally or popularly represented symbolically (e.g. discourse state, semantic meaning) while other knowledge is normally represented stochastically (acoustic confidence, n-gram language models). Eack knowledge source's reliability varies widely as different phenomena are being processed. For example, prepositions are poorly discriminated while most proper nouns can be accurately matched from an acoustic representation. If knowledge sources can be combined in such a way as to pay most attention to those things each knowledge source does well, discounting the information a knowledge source does poorly, we will be able to develop an overall more powerful metric from multiple knowledge sources. A metric that combines information from many knowledge sources will better determine whether a word is misrecognized.
- **Methods for inferring the meaning of novel words and phrases that are robust in the face of uncertainty, noise and inaccuracies.** It is not enough to detect an out-of-vocabulary word. To enable re-recognition and accurate post-processing it is necessary to determine how the word is used, its basic semantic category and its specific meaning. For example, if ***"the bay area"*** is a novel word string encountered while processing a dialog, the goal is to determine that it is a location, functions as an origin or destination in the domain, and refers to the San Francisco area. Unfortunately, it is not sufficient to merely adapt the methods used for acquiring novel words from text. Spoken input does not have the syntactic regularities of written text and is complicated by misrecognitions, ill-formed input, etc. On the other hand, spoken input is usually part of a larger dialog, enabling discourse structure and constraints inferred from prior, applicable context to be exploited when determining the meaning of a novel word or phrase.

3 Proposed Areas of Investigation

The goal of the proposed research is to enable spoken language systems to automatically acquire new words and learn from their recognition errors. Our research agenda for developing, testing and refining the necessary algorithms involves answering the following research questions:

1. **How can a system detect recognition errors?** How can we model unknown words and develop confidence measures for hypothesized words?
2. **How can we determine the reason for recognition failure and detect and delimit unknown words?** When should the system hypothesize a new word?
3. **How can we determine the meaning of a new word?** What semantic category or word calss does the new word belong in? What is its meaning? How is it used in the grammar? When there is not enough evidence to determine meaning, how to we maintain and refine hypotheses?
4. **How can we incrementally update the system lexicon and grammars for future recognition and understanding?** How can we update a statistical language model without corrupting its estimates? How can we reliably estimate how a new word will

appear in later input?

5. **How can we refine word meanings and system grammars as additional evidence becomes available from further usage?**
6. **How do we acoustically represent new words so they can be recognized?** What methods should we develop for robustly representing the new words, both for immediate usage in the dialog where they are first introduced and for later usage when spoken by different speakers in different contexts?
7. **How do we update existing lexical representations, language model and grammatical representations when errors are caused by the respective knowledge source?**
8. **How can we refine the lexical representations of new words?** Can we examine similar entries that mean the same thing and determine when they represent the same word?

4 Background and Pilot Feasibility

To understand the technical challenges posed by out-of-vocabulary words, we present the following background on the speech recognition process. This is followed by a description of related work in automatic word learning approaches in text and pilot work demonstrating feasibility of the proposed technical approach.

4.1 Speech Recognition

Today, the most widely used and successful continuous speech recognition systems are based on hidden Markov models (HMM's) (Bahl, Jelinek & Mercer, 1983; Lee, Hon & Reddy, 1990; Lee, 1989; Huang, 1991). Words are represented as sequences of phones (which may be context specific). The acoustics of each phone is modeled by a hidden Markov model. These models are trained from the series of feature vectors computed from a large set of training utterances. An acoustic model for a word is formed by concatenating the HMM's for the phones comprising the word. In order to decode speech, the word models are matched against the speech input. Most speech recognizers operate in a left to right fashion matching representations of acoustic sequences with ***only*** words contained in its lexicon, including silence and noise models (Ward, 1989; Wilpon, 1990). ***ALL*** parts of the input must be matched against a known lexical item. The search initially transits to the beginning of each word model. It then proceeds in a left to right fashion matching word models against the input. When a hypothesis comes to the final state in a word, it again transits to the beginning of each word model in the lexicon. Thus, in principal, the search is trying to match the acoustics for every possible sequence of words. The two most prominent search strategies are stack decoding, using an A* search, (Nilsson, 1969, 1971) and viterbi beam search, a dynamic programming algorithm, (Viterbi, 1967). In spite of these strategies, even a small lexicon has a very large search space. In order to make the search more tractable, language models are used to reduce the word sequences that must be searched for.

Language models may specify legal sequences of words or assign probabilities to sequences of words. The speech recognizers produce a string of words, guided by their language models. When the recognition search comes to the end of a word, it uses a language model to determine what words to search for following the current word. This predictive component greatly reduces the search space (Lowerre et. al., 1980; Erman et. al., 1980; Kimball, et. al., 1986).

Given the way speech recognition systems operate, it is easy to see how an out-of-vocabulary word disrupts the process. Recognizers only try to match known words from their lexicons against incoming acoustics. Since unknown words are not represented, they cannot be matched. Instead, unknown words cause substitution errors, some word from the lexicon is substituted for the unknown word. Further, the substituted word may not be the same length as the unknown word, causing the next word to be searched for in the wrong place. Also, since an incorrect word was hypothesized, the language model will most likely not give the correct set of next words to search for. This causes search misalignment because the grammar is no longer predicting the correct words and the search is no longer correctly aligned on word boundaries.

When this type of error is made, neither the system nor the user realizes the cause of the error. The system knows only that the total path score was poor. The user knows only that the system did not get the correct answer. Using more restrictive grammars reduces the search space more, but is also disrupted more by the unknown words.

4.1.1 Bottom-Up Approaches

Simply giving up language models and continuity constraints during the search has not worked well. Speech recognizers have been built which use control strategies other than a left to right word search. These approaches are appealing because they do not count on decoding previous words correctly in order to recognize a word. One technique is to look for each word in the lexicon starting at every point in the input and to generate a lattice of hypotheses that score better than a preset threshold (Adams and Bisiani, 1986; Chow and Roukos, 1989). This approach generally performs much more poorly than the full path score approach. The lack of constraint in the search yields word lattices of poor quality. Correct words will be missing from the lattice and there are many "false alarms" (word hypotheses with good scores, but the word was not spoken). There are also additional problems of normalizing and comparing segments of different lengths and of determining allowable junctures between the word hypotheses. (Stern et. al., 1987; Ward et. al., 1988)

Another approach is to decode the speech input into a string of phonemes and then parse the phonemes into words (Levinson and Ljolje, 1989; Levinson et. al., 1989). Again, due to lack of constraint in the search, the phoneme strings that are generated are of too poor a quality for the system to perform as well as the top-down systems.

Reducing constraints in these ways may increase the robustness of a system to unexpected input, but reduces overall performance of the systems. Our challenge is to be flexible to the occurrence of new words without giving up so much constraint that the overall system performance is seriously degraded.

4.2 Prior Work in Learning New Words: Feasibility

There are two lines of research that bear directly on the issue of how to automatially detect and acquire new words. The first comes from work in text processing and addresses methods for determining the meaning of the new words. The second consists of three pilot studies that demonstrate the feasibility of our approach. These investigations illustrate how well out-of-vocabulary words and, more generally, recognition errors can be automatically detected in (spontaneous) speech. Two of the studies demonstrate and perform preliminary evaluations on newly developed acoustic normalization techniques that permit us to measure confidence of the acoustic hypotheses.

4.2.1 Learning New Words from Text

There has been a good deal of research in learning new words from text. This is a much easier task than learning words from speech because there is no question whether or not an unknown word is present and text obeys syntactic rules that can be used to determine the role of the new word in a sentence. However, it is useful to review the techniques to see which if any can be adapted to spoken language word acquisition.

There have been four approaches to learning the meanings of new words encountered in text. The first uses scripts or representations of stereotyped action sequences to assign meaning to newly encountered words (Granger, 1977). The idea is to identify which part or action in the script is associated with the new word and to use this to extrapolate its semantic associations. While this is a powerful technique, there is a problem with the number and generality of scripts necessary to be useful. The second approach to new word acquisition in text relies upon user interaction to determine the syntactic and semantic roles of the new words. Originally introduced in the VOX system (Meyers 1985), the technique is popular (e.g., TEAM) and has been implemented using many varied architectures and knowledge sources. It allows the user to extend the vocabulary, events and scenarios a system can understand by asking a series of directed questions. For example, in entering a new word that is a noun, the system will ask for singluar and plural forms, synonyms, and often a parent concept or semantic case role. While this approach ensures information is input correctly and unambiguously, it is not really automatic.

The third method takes an incremental approach towards determining new word meaning. These systems rely upon multiple knowledge sources and preceeding information to generate candidate meaning hypotheses. When unique meanings cannot be derived, alternative hypotheses are maintained until further evidence enables a unique meaning to be inferred. Originally introduced in the POLITICS system (Carbonell, 1978), the idea was to use sentence context in conjunction with constraints derived from inferring the goals and plans of the speaker to hypothesize the semantics of a new word. Later incarnations (Jacobs & Zernik, 1988) relied upon different knowledge sources (e.g., the parse from the rest of the sentence, morphology, syntax, semantics, and context) but refined the incremental hypothesis refinement and convergence techniques.

The fourth approach exemplifies why most text-based approaches are inapplicable to spoken language. These methods rely heavily on syntax to constrain the meaning of a new word. For example, consider Lytinen’s (1991) work. His system computes the syntactic category of an out-of-vocabulary word as a pre-requisite to inferring meaning. When it can compute a new word’s syntactic category from grammatical knowledge (sometimes) it attempts to infer the semantic relation between the new word and other parts of the sentence by relying upon a standard hierarchical knowledge base.

There are two important lessons that can be applied from the text-based research. First, the successful systems use partial knowledge derived from multiple sources, including semantics, to infer word meanings. The trend is the more knowledge sources employed, the more successful the system. Although a spoken language system cannot use many of the knowledge sources available to a text system (e.g., syntax), it can use others from dialog. The second lesson is that it is sometimes necessary to maintain sets or spaces of alternative meaning hypotheses and incrementally refine these using future examples. Our approach incorporates both of these lessons, merging partial interpretations derived from many knowledge sources to infer word meanings and relying upon incremental refinement methods when prior context is insufficient to derive word meanings.

4.2.2 Pilot Study I: Detecting New Words in Speech

The first experiment on the feasibility of detecting out-of-vocabulary words in speech was conducted by Asadi et. al. (1991). The idea was since unknown words are composed of novel sequences of known, modelled sounds, an unknown word model could be developed. The model matched any sequence of 2 or more context-independent phonemes. The unknown word model competed with known words in seven pre-selected "open word classes" from the class n-gram language model (e.g., port_name, ship_name, etc). When a new word was detected (i.e., the "unknown word" model scored higher than competing known words in a standard recognition search) the system asked the user to type the word. It then looked up the word in a large phonetic dictionary to generate a word model from the phonetic spelling. To add the word to its grammar, the system displayed the set of seven "open classes" to the user and asked the user to pick the appropriate category. The word was then added to that class in the class n-gram language model. Read speech from the ARPA Resource Management corpus were used in the study.

This experiment showed the potential of using an explicit all-phone model in the recognition search and illustrated how to augment a class n-gram language model when a new word is encountered. Unfortunately, there was a major flaw in the study, as their "new words" consisted of previously (well) trained, known words that had been temporarily removed from their system’s lexicon.

Our work extends this approach in several important ways:

- Our all-phone, unknown word models use triphone transition probabilities to enhance the recognition accuracy in the same manner that a statistical language model enhances recognition accuracy in normal recognition. This approach is language independent, as the triphone transition probability matrix is trained for a given target language (English).
- We permit new words to occur in any "class" of words and any semantic category.
- We use the all-phone models as a means of normalizing acoustic recognition scores. This enables us to assess goodness of acoustic match independently of competing hypotheses, as described in Pilot Studies II and III.
- Higher-level knowledge sources are used to assist in detecting recognition errors and to determine the meaning of a new word.
- New words are ***automatically*** added to the system class n-gram language model as well as to the system's parsing grammar, semantics and lexicon. User interaction is not required.
- Acoustic representations of new words (word models) are automatically generated using a variety of experimental techniques.

4.2.3 Pilot Study II: Normalizing Acoustic Scores

This study illustrates the all-phone acoustic normalization procedure proposed here and a preliminary investigation of how to use normalized scores to assign confidence measures to recognized words.

Continuous speech recognition systems map known lexical entries onto all parts of the input. When a novel word is encountered, it is misrecognized and usually causes misalignments that cause recognition failures in the surrounding regions of input. To detect misrecognized words in speaker independent, continuous, spontaneous speech, we used a technology based upon a generic unknown word model that permits any triphone (context-dependent phone model) to follow any other triphone given n-gram phone transition probabilities. Vocabulary independent English triphones were used. Triphone transition probabilities were computed for triphone bigrams trained on a 20 million word corpus from the *Wall Street Journal*. Because our goals are to detect all recognition errors and to permit new words to occur in any grammatical class, this pilot experiment (Young and Ward, 1993a) computed confidence measures for all recognized words. To do this, we first normalized acoustic word scores using the type of all-phone models proposed. An all-phone string was decoded in parallel with the word decoding, as illustrated in Figure 1. The acoustic score for each word match was normalized using the score for the corresponding region of the phone decoding. Specifically, for each word we compute an acoustics-only score and subtracted from this the acoustics only all-phone score for the corresponding input frames. The result is the normalized score. Unlike the scores output by the recognizer that are useful only in comparing alternative hypotheses, the normalized scores provide a means of independently assessing overall goodness of match. In effect they estimate the prior probability of the acoustics unconstrained by word or word-sequence models.

To turn the normalized scores into a confidence measure, we used a Bayesian updating method, trying to estimate the probability that a word is correct when it has a given score. For this study we clustered words by similarity of pronunciation, and collected non-paramtric statistics for each group of words from a training set of 1000 utterances from ARPA's ATIS2 corpus. We came up with 153 word classes. The results indicated that while we weren't always able to accept or reject word hypotheses based on acoustic match, we knew how reliable the decision would be. This is the first step toward combining knowledge sources to make the final decision.

4.2.4 Pilot Study III: Detecting Errors in Spontaneous Speech

This study extends the approach outline in Pilot Study II, but used a different word grouping method for estimating confidence of recognition. Again, a phone-based decoding was run in parallel with the word-based search and recognition scores were normalized. The phone search provides an estimate of the acoustic match of phone models to the input unconstrained by the lexicon and the language model. In this study (Young and Ward, 1993a) we conducted two experiments to evaluate the utility of the normalization technique and to evaluate ability to correctly reject misrecognized words. We used an 1800 word lexicon that included 10 "noise words" to model filled pauses, stutters, partial words and environmental noise found in spontaneous speech [Ward, 1989]. The SPHINX-I discrete HMM-based speech recognizer (Lee et. al.,1990) and a bigram language model with perplexity 55 was used for recognition. SPHINX-I outputs a single best word string for each recognized utterance. Spontaneous spoken utterances from ARPA's ATIS corpus were used in these experiments. These utterances were typical of spontaneous input, containing both the acoustic (filled paused, stutters, etc.) and structural (ill-formed, restarts, mid-utterance corrections, etc.) phenomena normally present.

The first experiment assessed whether the normalization procedure provides a more useful score that the relative scores normally output by a recognizer. Further, this experiment assessed our ability to correctly reject misrecognized words for each of the 1800 words in the lexicon, ignoring the effects of word frequency. We generated sentence hypotheses for 5000 ATIS utterances using SPHINX-I. For each word hypothesized by the recognizer, we stored its acoustic word scores (all-phone and regular score) and a flag indicating whether the word was correctly recognized. From this data we created signal (correct) and noise (incorrect) distributions for each word. We assessed the system's ability to correctly reject misrecognitions looking at the measures of d-prime and power. D-prime measures the difference between the means of the signal and noise distributions. The larger the d-prime, the greater our ability to correctly reject misrecognitions. Similarly, power assesses ability to correctly reject misrecognitions at a given "miss level" where correctly recognized words are rejected. We defined the measure ***power*** to be the percentage of incorrect hypotheses that will be rejected for a cutoff that would only reject 5% of the correct hypotheses.

Before normalizing, the average power for the 1800 words in the lexicon was 65% based upon using acoustic scores. After normalizing, the average power increased from 65% to 74% based upon the 5000 utterance test set. This indicates that, in general, the normalization procedure makes correct and incorrect words more separable.

The second experiment assessed our ability to correctly reject misrecognized words when normalized scores are turned into confidence measures. We attempted to estimate the probability that a word is correct when it has a given, normalized score. Words were clustered by using their signal and noise distributions. For each class of words, we quantized the range of scores into 75 bins or score ranges. We then took normalized word scores from 5000 utterance recognition hypotheses (~30000 words) and accumulated histograms for each word class. For each bin in a word class we determined the percentage of the time words in the class with a score in the bin were correct. These histograms were then smoothed. This gives us a direct measure of confidence that a word is correct when it has a given acoustic score. For this experiment, we used a test set of 1000 spontaneous utterances from the ARPA Feb92 ATIS test set. The test set contains words never seen in training and the results reflect our ability to correctly reject misrecognitions while taking into account word frequency effects in the test set. These word frequency effects are what distinguish this experiment from Experiment 1. Again, we set a rejection criteria to maintain 95% correct accepts and measured ability to reject misrecognitions. For this test set, the correct acceptance rate was 94% and the rejection of misrecognized words was 53%. In other words, we could accurately detect 53% of all misrecognized words in the 1000 utterance test set while at the same time only rejecting 6% of the correct words. In looking at the histograms for the word classes, some had almost perfect classification, while others had only slightly better than chance. Function words (high occurence) were most poorly rejected. However, discrimnability still varied. For some word classes, we can very reliably accept correct words and reject misrecognitions on acoustic evidence.

5 General Technical Approach

Detection of new words in speech input is a difficult task. It increases ambiguity in the search of the speech recognition process. Our approach to learning new words from speech is based upon more general methods that enable a system to detect and correct its errors. The approach balances constraint against flexibility by capitalizing upon the multiple knowledge sources available to speech understanding systems. A speech understanding system has many available sources of knowledge: a lexicon, stochastic and rule-based grammars, semantics, acoustic models, domain knowledge, dialog structure and constraints from prior interaction. This research develops methods to enable the conjoined use of all available knowledge for detecting errors and determining how to correct them. The research will be developed and demonstrated using speaker-independent, continuous, spontaneous speech and the techniques will be applicable minimally to all Hidden Markov Model based understanding systems.

To detect misrecognized words, we will develop confidence measures that take into account acoustic goodness of recognition and discourse based information derived by trying to understand the recognized string. Once we have determined that an area of input is poorly matched, we will differentiate the presence of new words from the other problems that can cause poor recognition (e.g., filled pauses, stuttering, partial words, language model violations, inadequate grammatical coverage and alternate word senses.) We will develop a variety of tests to determine the cause(s) of an error.

In order to be included in future searches, new words must be added to the system's language model. Our approach to deciding how to include the newly acquired word in the system's language model is based upon first determining candidate word meanings using the dialog, semantic and pragmatic components of our existing system. We draw upon the lessons from textual word learning and merge partial information from many knowledge sources to infer word meaning. Further, when insufficient information is available to make a single meaning determination, we maintain a set of hypotheses that are incrementally refined with further input. In such cases, we initially add the word to multiple grammatical categories and later refine the language model and parsing grammars. Since the language model is a class n-gram and the parsing grammar is composed of nets indexed to concepts, if word meaning can be inferred, we know where to incorporate the newly acquired words. It is particularly important to accurately add new words to a system's language model. If words are too strongly predicted, a system can "hallucinate" them when they are not present (Reddy and Newell, 1974). Similarly, words cannot be matched if the system does not search for them.

Adding the word to the lexicon requires developing a robust representation of the new word string. Our approach requires that we develop an initial lexical entry for hypothesized new words for immediate use in the following utterances, as speakers may utter a new word more than once in a dialog. After the dialog is complete, these representations will be refined and generalized. We will experiment with a variety of methods to generate robust lexical entries for both new and commonly misrecognized words. One method is to use word spelling. Recent experiments (Alleva, 1989) have shown you can spell a word from only its acoustic match with a reasonable degree of accuracy. Similarly, we can experiment with acoustic generalization techniques including merging the matched sequences output by different types of recognizers (HMM vs. neural net). Our general approach to the potential problem of overgeneration of word models is to include a stage in later processing that looks for similar word models within semantic word categories. If two words map to the same semantic concept and have similar lexical representations, we will consider merging them.

Finally, we will try to correct errors caused by other knowledge sources, extending and modifying them, where appropriate. We will adapt the techniques outlined for adding information about new words.

In the sections that follow, we give an overview of our system architecture and then we outline our methods and illustrate them when possible using the following examples (unknown words are italicized).

REF: ***WHERE THE HELL*** **is ******* stapleton**
HYP: WHAT * **** is DALLAS' stapleton**

REF: please list all single engine ***OR DOUBLE*** **engine planes from pittsburgh to baltimore on august fifth**
HYP: please list all single engine ARE AVAILABLE engine planes from pittsburgh to baltimore on august fifth

```
REF: show me the TELEPHONE numbers of the car rental agencies
     in washington d c
HYP: show me the TELL OF ALL numbers of the car rental agencies
     in washington d c
```

These example illustrate the three types of out-of-vocabulary words and phrases encountered in our training data. They are interjections, new ways of saying known concepts and words that express contextually appropriate ideas that exceed the application domain, respectively. The system will try to model and add all of the above new words. However, interjections and words associated with concepts that exceed the application will not be mapped into database query commands. Also, the system is designed so that extra-domain concepts known by the semantic module will be processed normally. However, the system will not be able to accurately derive a single meaning for those concepts not represented in the domain semantics. In the sections that follow we will illustrate system processing using the second example.

5.1 Proposed Architecture and System

We propose to extend the existing Carnegie Mellon Spoken Language Understanding System for this research. Our current system has a loosely coupled architecture consisting of the Sphinx-II speech recognition system, the Phoenix robust speech parser and the Minds-II knowledge-based postprocessor. In this system, the initial recognition search is guided by a class bigram grammar. The recognized string is then parsed using the Phoenix (Ward, 1990, 1991) caseframe parser which has been designed for robust processing of spontaneous speech. Phoenix uses semantic grammars compiled into Recursive Transition Networks (RTNs) to specify word strings corresponding to concepts, and associates concepts in caseframes. The RTNs for concepts are semantic fragments. For example, the utterance "I want to see flights from Boston to denver after 5 pm" would be the concept sequence [list] [select_field] [from_location] [to_location] [depart_time_range]. The pattern matcher looks for phrases that map into concepts and then tries to put as may of these together as possible. The parser generates a set of interpretations that account for as much of the input utterance as possible. It skips regions and produces partial interpretations if it cannot produce a complete one. The conceptual analysis postprocessor Minds-II (Young, 1993e; Young & Matessa, 1991, Young and Ward, 1993c) analyzes hypothesized word strings and parsed output to detect misrecognitions and correct inaccurate parses. When recognition errors are detected, the system generates a set of meaning hypotheses for the misrecognized region, translates these into RTN networks and sends the region and hypotheses to the RTN decoder (Ward and Young, 1993) for re-recogntion. Results evaluating performance on 1250 ARPA ATIS utterances indicate the system can detect 98% of all semantically inconsistent recognition errors, generates correct predictions for 88% of these and corrects 50%. The system detects between 60% and 75% of all recognition errors. It cannot detect semantically consistent errors such as:

I would like information on ***ONE WAY*** *flights from Boston to Pitt*

I would like information on ***MONDAY*** *flights from Boston to Pitt*

There are two important aspects of this system with respect to the proposed research: its ability to detect recognition errors and its ability to generate meaning hypotheses for unknown or misrecognized input. The system keeps track of all preceeding interaction (user input, system responses and changes in screen displays), if any. The procedures for detecting errors and generating predictions are similar and both rely upon generating constraints that the utterance must fulfil by using domain semantics, pragmatics, previously communicated information and a domain independent discourse model. The discourse model determines the types of subdialogs that can be initiated at a given point in time and the domain topics that are current and can be pursued at each point in the interaction.

To detect errors, the system looks for semantic or pragmatic inconsistencies in the recognized string. Within an utterance, each of the phrases and words must modify or complement one another. Across utterances, contextual constraints on what is referenced, what subjects are available for discussion, and what types of subdialogs can be initiated must be adhered to. Plan tree traversal heuristics [23] that indicate what topics are available for discussion and requisite ordering among topics or plan steps must not be violated. To detect recognition errors, the system uses abductive consistency to determine how to best interpret and combine information to form a meaningful utterance. To determine which portion of an utterance (if any) is most likely to be inaccurately recognized, the analysis routines try to build the best, most encompassing semantically and pragmatically consistent representation of an utterance. It takes into account heuristics for processing restarts and mid-utterance corrections and tries to build a single semantic representation of an utterance, identifying the least number of semantic objects or attributes that are inconsistent.

Once the system has identified a misrecognition, it tries to correct it by by generating meanings for the misrecognized region. Here the system searches the knowledge base for the most general, contextually appropriate concepts that satisfy the constraints upon the utterance and are consistent with the other concepts in the utterance.

The predictions are translated into semantic nets and constraints upon them (as illustrated in Figure) and used to guide re-recognition. Re-recognition is performed by a speech recognition system that uses Recursive Transition Networks as a language model to control word sequences searched for when decoding an utterance (Ward, 1993). The recognizer uses the RTNs used by the Phoenix parser. . This recognizer uses discrete-HMM's to represent context-dependent phone models and is based upon the Sphinx-I system (Lee, 1989). In this version, the word transitions in the decoding search are guided by the Recursive Transition Networks used by the parser. The system searches for a sequence of concepts, where word sequences constituting concepts are specified by RTNs in the regions specified by the conceptual analyzer.

Figure 1: Dynamically Derived Net Restrictions

In our proposed system, we will use use a system similar to the above one. However, we will incorporate our unknown word model and all-phone decoding into the system and any and all new techniques developed. Additionally, we will adapt the Sphinx-II system for recognition. Our current Sphinx-II system has considerably better acoustic modeling than Sphinx-I. It also has an improved search algorithm. We will need to develop a Recursive Transition Network search version of Sphinx-II for the project. Sphinx-II will produce an initial decoding of the speech using a class bigram grammar. In addition it will decode an all phone beam in parallel. The word string produced by the recognizer will be passed to Phoenix to be parsed. The Phoenix output will be analyzed by a mechanism similar to MINDS-II to detect portions of a hypothesis which are likely to be incorrect. An acoustic confidence for each parsed concept will also be produced by the technique outlined in (Young, 1993) . Once a portion of a recognition hypothesis is marked as possibly misrecognized, hypotheses regarding its content will be generated and we will begin to determine whether an unknown word is likely to be present. One of the techniques we will use to determine whether a new word is present is to reprocess the misrecognized input using an RTN decoder, the all phone model and predictions derived from a Minds-II like component. Further, the meaning hypotheses generated for the misrecognized region will become the initial hypothesized meaning(s) of a new word, should we decide one is likely to be present in the region.

5.2 Alternate Architecture: Single Pass

The central issue for this research is the combined use of multiple knowledge sources to make decisions about the content of an utterance. In particular, how these sources can be optimally combined in specific situations to give the best overall decision. The search architecture we have described is one in which some knowledge sources are used to form an initial hypothesis which is then followed by analysis and reprocessing using additional knowledge sources. Some of the most useful information comes from analyzing the overall consistency of an utterance. The proposed architecture is convienient and efficient for this sort of analysis. However, other architectures can also be used which may have other advantages. In particular we will also build a system in which all knowledge is applied during the initial recognition search. For this we will use the same knowledge sources used in the multi-pass architecture but apply them in a predictive rather than analytical fashion. At CMU, we have already conducted some initial experiments with this sort of system. First, we developed the original MINDS system that constrained what could be recognized in an input utterance by restircting input to "what makes sense" given prior interaction, inferred user goals and plans and previously communicated information. Our experiments with MINDS illustrated that higher-level knowledge sources could be used predictively, to constrain candidates for recognition. While this system reduced perplexity of search in excess of an order of magnitude, there is a second way that higher-level knowledge could be applied during the recognition search, namely to prune and rank alternative word hypotheses. Furthermore, we have conducted experiments on developing predictive language models to directly model when a new word can appear. These experiments illustrate that it is possible to train language models to predict the likelihood of a new word and to produce probabilities for the words that can follow the unknown word using a stochastic language model. The class bigram language model represents the probabilities of seeing unknown words in classes as well as the standard set of known words. An unknown word is represented by an all phone model with bigram probabilities for phone transitions. This much is like our original proposal. However, now the higher level knowledge sources will be used predictively to reduce the overall search space by semantic content, including the new word search. This is essentially what is done in the reprocessing step in our original algorithm. The difference is that now all predictions are made before and during the original decoding pass, rather than as a result of analyzing the initail hypothesis. This will reduce the ambiguity of the search and therefore make the task of recognizing a new word as opposed to an incorrect sequence of known words easier. Our reliance on semantic information (in addition to syntactic) makes this scheme less likely to be confused by lack of coverage in the grammar of sequences of known words.

5.3 Modeling Unknown Words: Acoustic Normalization

One source of knowledge used for detection and recognition of new words is explicit acoustic models for out-of-vocabulary words. A system models each word that it knows as a specific sequence or network of phone models. By definition, an unknown word consists of a series of known, represented units, namely phones, but in a sequence the system does not search for. Our approach to modeling unknown words in a spoken utterance is to explicitly create a prototype "unknown word" model that allows any triphone to follow any other triphone (given context). Triphones are context dependent phone models. In the same way that stochastic models of word sequences (the language model) are used to control the recognizer search at the word level, we use stochastic sequences of phones to control the all-phone search. We will use trigrams of triphones to represent the probability of seeing sequences of triphones in a given target language (English). These sequences will be trained from a large body of text. We intend to use all freely available computer readable texts, including newspapers, novels and the Brown corpus. Such a probabilistic model represents the acoustic properties of the selected language without regards to particular words, word boundaries or grammar. These acoustic properties are not only important for initially modeling out-of-vocabulary words, they also enable acoustic word scores to be normalized independently of the other competing word hypotheses, as described in the pilot feasibility studies (Pilot Studies II and III).

5.4 Detecting Misrecognitions and New Words

As described in the feasibility studies (Pilot Studies II and III), we use the "unknown word"model, or all phone beam to decode speech in parallel with the word search (see Figure 1) to compute acoustic confidence measures. For each word hypothesized, we know how likely it is to be correct and how well we can determine if it is inaccurate.

Figure 2: Parallel Decoding for Normalizing Recognition Scores

For example, we know how well we can correctly accept and correctly reject each of the words in the recognized string ***please list all single engine ARE AVAILABLE engine planes from Pittsburgh to Baltimore on August fifth***. Our pilot data indicates that we correctly accept and correctly reject Pittsburgh 100% of the time. Baltimore, august and fifth are also discriminable. In contrast, are and available can be correctly accepted 100% of the time but cannot be correctly rejected at all. This means that if the normalized acoustic scores computed by subtracting the all phone score from the word score were such that we suspected Pittsburgh or Baltimore, we would be able to reliably reject it. However, the errors in ***"are available"*** cannot be rejected using acoustic difference scores. The point is that we know whether we can rely upon the normalized acoustic word scores to detect misrecognitions. Similarly, in the single pass architecture, we know both the reliability of recognized words and can compute the probability that an out-of-vocabulary word is present using the normalized acoustic scores in conjunction with higher-level knowledge as described in the next paragraph.

For those words which we can't discriminate on acoustic evidence, some other form of evidence must be used. We know from our previous work that semantic, pragmatic and discourse level constraints can reliably detect many misrecognitions. Hence, we will use our higher-level knowledge sources in conjunction with normalized acoustic scores to decide when words have been misrecognized. In this example, the semantics indicate the region "are available" or "are available engine" are likely to be wrong. Since we know the acoustics are unreliable, we will want to rely upon the semantic determination.

5.5 When to Hypothesize New Words

Having determined that an error is likely, the system then must determine its likely cause(s). To determine whether recognition errors were caused by out-of-vocabulary words, we will use a variety of techniques to rule out other sources of error. In this way, we avoid spurious "new word" hypotheses. There are many reasons a word may be misrecognized. Sources of error include novel acoustic patterns due to an out-of-vocabulary word, inadequate statistical language models, novel word senses, and generally poor acoustic discriminability, among others.

The proposed research will develop methods for distinguishing the most likely knowledge source responsible for a misrecognition. To begin, we will use acoustic knowledge regarding patterns that are poorly recognized as well as the following techniques:

- all word recognition within the "misrecognized area",
- prediction-based recognition within the "misrecognized area",
- prediction-based word spotting in the misrecognized regions.
- trained language models to predict presence of novel words and semantic categories of common out-of-vocabulary words encountered in the application domain.

For example, an error caused by inadequately training a statistical language model will not be made when the language model's constraints are removed, either by performing an all word search or a prediction based search on the misrecognized input. Further, the normalized acoustic word scores output by these alternate recognition methods would differ significantly from the initial score (properly normalized) if the error were caused by a language model problem. Similarly, a word that is consistently misrecognized in some dialogs but not in others may indicate that an alternate word sense has been discerned, particularly if the word is correctly recognized when an all-word language model is applied. Errors caused by out-of-vocabulary words will not benefit from an all word search or prediction-based word spotting. Words that can be correctly recognized using prediction-based re-recognition indicate the semantic RTN grammar has adequate coverage to process the input. On the other hand, if a word string can be recognized with prediction-based word spotting and not with prediction-based re-recognition, we have an indication that the grammar coverage is insufficient. In contrast, the all-word approach makes no assumptions. Information about all detected errors will be accumulated in meta-knowledge data structures so that inadequacies can be corrected automatically by extending the system grammar, lexicon, language model, etc. at a later point in time.

The idea here is to develop a series of consecutive assessment techniques to infer the most likely error source. When these tests indicate a new word is likely, the system will propose a new word and determine an initial set of meanings for it. It is important to realize that we are not trying to make the determination of a new word on acoustic evidence alone. This decision must be made using all sources of knowledge used for constraining the search. In this way, we avoid over-generating new words. By using higher-level constraints, the system will be able to detect a new word that is pronounced similarly to a known word if they are associated with different semantic categories. On the other hand, the system will not hypothesize a new word if within the same semantic category there is known word that closely (one phone difference) matches the acoustics. Only a very poor acoustic match for words with high language model probability (out-of-vocabulary word) or a very poor language model probability for words with a good acoustic match (new word sense) would trigger the new word model.[1]

5.6 Generating Meaning Hypotheses

When new words have been identified, the system determines their meaning and uses it to augment the statistical language model and the semantic RTN parsing grammar. The initial set of meaning hypotheses for a new word is determined by the semantic, pragmatic and discourse knowledge used to generate recognition and re-recognition predictions in Minds and Minds-II, respectively. It is unlikely that anything more than semantic class can be reliably inferred from a single utterance. Thus, the system uses the context of the entire surrounding dialog to determine the meaning of a new word or phrase, employing both forward and backward inferencing techniques.

When available knowledge (the information conveyed or inferred from the utterance containing the new word string and from prior interaction) does not support a unique meaning for a new

---

[1]In order to prevent noises from being interpreted as new words, we will include noise models in the decoding search. This has previously been done with very good (98% accuracy) results (Ward, 1989; Wilpon, 1990). The noise models include both user noise (filled pauses, breath, etc.) and environmental noise (phone rings, door slams, etc). The acoustic models for these are clustered to give better coverage of non-speech sounds. The models are vocabulary and task independent and do not need to be retrained for each task.

word or phrase, even if a single semantic category is inferred, multiple hypotheses will be maintained. These meaning hypotheses will be refined using abductive reasoning and by (re)analyzing any prior interaction. The most probable semantic categories and semantic bindings will be stored for later confirmation in a data structure. The system design allows for forward inferencing. The alternate meaning hypotheses are analyzed each time additional information becomes available later in the dialog or from a later dialog. In most cases, a single dialog will be sufficient to determine word meaning. To derive a unique meaning, we intend to associate probabilities with the possible meanings. This involves keeping track of the number of instances consistent with each usage. When support for some hypotheses falls clearly below that of others, they will be eliminated in favor of those with more support. When there is insufficient information to determine a unique semantic category for the new word(s), initially the word will be added to all candidate grammatical categories in the language model and grammar. This will enable the new word to be searched for and recognized in later utterances until a unique grammar category can be inferred.

An additional system attribute is that new words can have multiple, consistent meanings. For example, ***"the bay area"*** can serve as a city (San Francisco) and an area. Similarly, in the utterance ***"What ground transportation is available from Stapleton to BOULDER"***, Boulder could either be a city name or the name of an airport. We will not insist on the convergence into one meaning for a word. Rather, the criteria for convergence is that the alternatives either be semantically related or fill the same semantic case or role in an utterance (and are interchangeable). In the above example, both meanings for "Boulder" are legal fillers for a destination or an origin.

To illustrate, consider the example utterance

- **"Please list all single engine ARE AVAILABLE engine planes from Pittsburgh to Baltimore ..."**

where "are available" is substituted for "or double". The utterance constrains the region meaning to be an attribute of aircraft type and potentially an attribute of an aircraft engine. The preceding dialog supports this interpretation. Knowledge of mid-utterance corrections further supports the hypothesis that the utterance relates to aircraft engines. The resultant hypothesis set therefore will include "aircraft engine attribute" including single, double, turbo-prop. Since the application does not have information on turbo-prop planes, this alternative will be weighted less. The mid-utterance correction evidence supports both the single and double engine hypotheses. However, the dialog structure knowledge source is biased against repetitions unless they are part of a mid-utterance correction. The result is the "double engine" hypothesis has more support. However, both will be maintained and evaluated when evidence becomes available.

- ***"Show me the TELL OF ALL numbers of the car rental agencies in Washington D.C."***

The utterance above is processed similarly. However, in this case, prior interaction strongly indicates that ground transportation is being further discussed. The system knows that "numbers" and "car rental agencies" have all been reliably recognized. The system’s semantic component knows that an attribute of car rental agencies that involves numbers is being requested. There are only two attributes that can apply, street numbers and telephone numbers. So, even though this query exceeds the application, the system knows enough about businesses and car rental agencies to infer some meaning.

Many of the proposed techniques for generating candidate phrase meanings are implemented in the Minds-II system. The mechanisms underlying Minds-II abilities are constraint satisfaction techniques in conjunction with abductive reasoning and basic syntactic knowledge of constituents and attachment. These are general, domain independent techniques which rely upon a domain specific knowledge base. The system begins by hypothesizing which entities and actions in the utterance the new word or phrase could modify. This process relies primarily on syntactic knowledge. Next, the system hypothesizes reasonable semantic values that the region could take. This procedure uses abduction and constraint satisfaction. Semantic constraint violations include both type constraints of objects and attributes and n-tuple constraint violations.

To illustrate a semantic n-tuple constraint, consider the following rule for long range transportation:

Long Range Transportation

Objects:

vehicle long-range vehicle,
origin location,
destination location
objects-transported object

Relations:

origin - destination

The example illustrates type constraints on what objects may fill a slot and relational constraints. The relational or tuple constraint here restricts the relationship between the origin and destination slot fillers. The constraints have a twofold effect on deriving word meanings. First, new word meanings that violate a constraint are not generated. Second, the constraints are used to compute potential meanings. So, in the "rental car phone number" example, the constraint on what attribute of a business takes a value of "number" was used to narrow the possible meanings for an out-of-vocabulary word that referred to an out-of-domain concept! The set of possible meanings is further constrained and supported or refuted by applying information about the surrounding dialog and rules from the discourse model. Each source of knowledge (discourse structure, attributes of restarts and mid-utterance corrections, attributes of any applicable subdialog, domain tree traversal rules and any constraints propagated by prior interaction or what is available for reference) is used to further support or refute a hypothesized word meaning.

5.7 Adding New Words to a Statistical Grammar

Once new words have been detected in spoken input, it is necessary to associate them with grammatical categories and add them to the statistical grammar. In this system, association with the grammatical category has already been done by determining meaning. Even when a single semantic (grammar) category cannot be determined, the new word or phrase is added to the candidates. So, in the "engine" example above, the new word string would be added to the class of type of engine where "single" "jet" and "turbo-prop" would be stored if they existed. In the "Boulder" example, Boulder would be added to the city-name and airport-name classes.

To add the new word(s) to a grammatical category in the class n-gram language model, we use the following procedure. First, the language model considers all words within a class to be equally probable. The classes used in the N-gram will (and do) correspond to the meanings or categories in our semantic Recursive Transition Networks. We will expand the classes to words dynamically during the search, so we only have to add the new word to the list of words comprising the appropriate class. When the search reaches the end of a word model, it determines the probability of transiting to each class from the class of the current word. A transition is then made to each word in the class with the transition cost for each word being (1/N * class transition cost). If word transitions are precompiled instead of class transitions, the new word can still be added, but the procedure is more complicated.

An additional advantage of classes is that we can gather statistics on the frequency of occurrence of new words by class. In many tasks new words will occur in some classes much more frequently than others. For example, in an Airline Travel Information task, many new words are in the "city_name" class, reflecting lack of knowledge by the user of exactly what cities are contained in the database. This knowledge may be useful when multiple, competing meaning hypotheses cannot be resolved.

5.8 Augmenting the Phrase Grammars

The meaning of a new word or phrase and its semantic category are used to add new words to the RTN grammar that is used for parsing and to guide any prediction-based recognition. The system is structured so that most semantic concepts represented in the domain knowledge base (all those included in the application) are pre-indexed to RTN nets and to the semantic tokens in associated RTNs. These nets define the semantic and syntactic contexts where the new word can appear in later utterances.

We have already developed routines which allow us to add new word strings to a network grammar. The networks are compiled separately and may call one another. Since a concept is represented by a separate network (rather than being a part of one large network) it is easy to add new nodes. When a new word is identified as belonging to a particular concept ("city" for example) it is added to the rules for the concept. In order to add the word to the compiled version of the net, it is converted to a word number and a new node for it is created in the in-memory copy of the network. A transition to the new node is added from the start node of the net. This new compiled net is then written to a file, replacing the old compiled net. These operations are very fast since each individual net is small. The system is not restricted to adding single words. Word sequences can be added. For example "the bay area" can be added to the "city" concept.

We will experiment with alternate methods for incorporating new concepts into the RTN (and class n-gram) grammar. One method is to look for other nets that modify the same semantic concept. In the example used, telephone numbers are a new concept. However, the system knows about "numbers" and "rental car companies", the two concepts modified. So, the new concept would be used to generate a new frame slot in the associated concepts. The new words associated with the new concept can either be added to the net associated with a similar slot filler, or a new net can be generated that reflects the input utterance. In such cases, the system will respond by telling the speaker they have exceeded the domain and by presenting the information that is available on rental cars. However, the new words do need to be added to the grammar and language model so they can be recognized if spoken again.

5.9 Augmenting the Lexicon: Initial and Final Representations

New words and phrases must be added to the system lexicon so they can be recognized if spoken again. The system is designed to add an initial representation of the word to the lexicon immediately following recognition. Later, a more robust, generalized model will be developed and will replace the initial (new) word model. The new word models will be represented in the same manner as all other known words in the system lexicon. In other words, the word will be represented as a flat sequence of triphones, each associated with an existing HMM.

We will investigate alternate methods for generating robust word models. Initially, we can use the sequence of phonetic units (triphones) matched by the all-phone recognizer when the unknown word was hypothesized. As explained above, an item which is close in pronunciation to a known word will only be added as new if its language usage (or meaning) is different from acoustically similar known words. It is not important that the pronunciation be exact, that is, the same as would be assigned to it by a linguist. It is sufficient that it be close enough that the word will be recognized correctly in the future.

To derive a more robust lexical entry for a new word, we will investigate the following techniques (among others):

- Merging and generalizing the sequence of phonetic units matched by the all-phone recognizer when multiple instances of the word were encountered.
- Merging or generalizing the sequence of phonetic units matched by basically different recognition techniques (i.e., neural net, semi-continuous HMM, discrete HMM).
- Generating an orthographic transcription and attempting to generalize it.
- Building new word models directly from senones rather than triphones.

What ever method results in the most robust, accurate recognition of the new word or phrase will be used. Thus, all methods will be empirically evaluated. Similarly, we will store all the lexical representations that can be generated for a new word, adding new instances each time a new word is encountered in later input.

To spell new words and phrases, we will need to know word boundaries. Word boundaries are already delimited as those frames which are matched to the unknown word model. The word is also already assigned a grammatical classification. In order to spell the word, we will use use HMM letter models similar to phone models. These models can be specific to the grammar categories. For example, proper names may have different models than action verbs. Preliminary versions of such a spelling system have been developed at CMU (Alleva, 1989). In some applications, correct spelling is more important than others. In dictation, the task is to spell the words correctly. However, for a front end to an application, it is possible that the system can have the correct semantic mapping and take the correct action even though the word is not spelled correctly.

5.9.1 Evaluating the Lexicon

Once new words have been added to the lexicon, we will evaluate it to determine whether any words are represented multiple times and to eliminate any falsely hypothesized new words. To do this, we will examine the word models associated with each semantic category or class separately. The implications of this are that words with multiple senses can have a lexical representation associated with each of its senses. Also, similar words can be represented if they have different meanings. However, those words that mean the same thing, especially those that have the same semantic binding, not just the same semantic category, and are similar will become candidates for merging. Since we will have stored information about each time a new word appeared, we will be able to use this evidence to determine the likelihood that the candidates for merging are the same word. Also, we can use the stored information to assist in deriving the merged or generalized lexical entry.

5.10 Correcting Errors in Other Knowledge Sources

Our approach to detecting new words involves detecting insufficiencies and inadequacies in other knowledge sources as well. Specifically, we will be able to detect insufficient training in language model, inadequate grammatical coverage in the RTN grammar and poorly trained word models. To modify and update these knwoeldge sources we will adapt the techniques used for adding new words to these knowledge sources.

5.11 Comparison of Approaches

The goal of this research is to enable speech recognition and understanding application to recognize and understand new words and phrases when they are encountered in spoken input. We seek to develop robust and accurate techniques that are domain independent and do not require profuse amounts of application specific training data to be reliable. Prior researchers have attempted to develop generic out-of-vocabulary word models but have not met with much success automatically detecting these words. A second, common approach is to require user interaction to add new words to an application. Our approach is designed to be automatic. We have elaborated upon the generic word modelling approach by developing an acoustic normalization technique. This technique has proven to be empirically robust and is most promising because it also provides a precise, independent assessment of recognition accuracy. The side effect of this technique is that it provides us with a method for combining other knowledge sources with the acoustics and enables us to combine knowledge sources based upon their reliability. This acoustic normalization technique is the basis of our out-of-vocabulary word model and the conjoined use of acoustic normalization, semantics, pragmatics, discourse properties and predictions nad language models form the basis of our approach.

We propose to investigate three complementary techniques for acquiring out-of-vocabulary words: single pass, multi-pass and direct training of knowledge sources. Each of these approaches has strengths and weaknesses, as described below. However, by adopting a comprehensive approach we will be able to empirically evaluate each approach and develop comprehensive and robust techniques for acquiring new words when spoken in spontaneous, speaker independent, continuous spoken speech.

The multi-pass approach has the advantage of building upon existing systems. The technique detects all errors, not just misrecognitions caused by out-of-vocabulary words. Because the system analyzes input incrementally, it does not have to contend with multiple competing hypotheses complicating the recognition search. The recognition search mechanism is known to be robust and operates in close to real time. The single pass approach more closely approximates human speech understanding. Like the multi-pass approach, new words are recognized in part by determining what they must mean. Unlike the multi-pass approach, all knowledge sources are applied during the recognition search, although both approaches use the same knowledge sources. The single pass approach has the potential drawback of requiring much more time. More significantly, this approach must answer questions about how to best apply constraints and predictions from higher-level knowledge sources during the search process. We know that when too many hypotheses are considered at once, systems often prune an accurate interpretation in favor of an inaccurate one before enough evidence can be accumulated to clearly prefer one or more of the hypotheses. The third technique involves directly training knowledge sources to recognize unknown words. Successfully trained and robust predictors of new words can be incorporated into either of the other two architectures. Prior experience indicates that training can often enhance any knowledge source. The power of any single knowledge source can be empirically evaluated. Candidates for direct training include language models, semantic categories and acoustic models. The drawback to this approach is that it may not generalize across application domains and may require profuse amounts of application data to be adapted to any specific application. On the other hand, stochastic models often capture many diverse knowledge sources and can be simple to train.

Viewed differently, we intend to pursue a comprehensive research program to automatically acquire out-of-vocabulary words as they as spoken. We will investigate two search mechanisms, one that constrains recognition and one that detects recognition errors by applying higher-level knowledge sources and combining them with acoustic normalization techniques. Any knowledge sources that can be effectively trained will be incorporated in the final system.

6 Performance Evaluation

6.1 Data

To train triphone transition models we will use large on-line bodies of English text such as the Wall Street Journal, Brown corpus, novels and newspapers. Triphone transition sequences can be automatically generated from orthographic transcriptions using existing pronunciation dictionaries. The text provides many, varied sequences of words, most of which are not included in any domain specific spoken language system's lexicon. A large dictionary of pronunciations is used to turn word sequences into triphone sequences. This results in a large number of triphone sequences to train our stochastic model.

We will evaluate the performance of the system on spontaneous spoken input from the Air Travel Information Service (ATIS) task. We also propose to use Scheduling as an optional second task for evaluating the system. The advantage of these is that they are existing data, so we do not have the added expense of gathering such data. Also, since other sites use the same data, the results will be able to be more meaningfully compared to research by other groups. The data gathered for these tasks is of the appropriate type, spontaneous speech from users engaged in performing a task. ATIS is a database query task in which a human interacts with a computer, while Scheduling is a translation task where two humans are interacting. For both tasks, we will have well defined training and test sets. We will develop systems based on the training data and evaluate them on the test data. Subjects in the test sets were not used for training data. We will not artificially remove entries from our lexicon to test the system. This does not give an accurate test, either for acoustic and grammatical coverage or frequency of occurance. We will use the ATIS2 training data for training and the ATIS2 test sets and all of the ATIS3 data for testing. ATIS2 is a corpus of spontaneous speech gathered from users performing tasks related to air travel planning. There are approximately 12000 utterances of training data and several test sets ranging from 300 to 1000 utterances per set. We already have a system which has reasonable coverage on the ATIS2 data, but ATIS3 represents an expanded database so should provide a good (and natural) test for new words. This is a good test of the ability of the system to extend itself to adapt to a database expansion. The ATIS2 data is for a 10 city database. For the ATIS3 data, the database was expanded to include 46 cities and all additional flights, airlines, etc. associated with them. We have not yet trained the system on any ATIS3 data.

We currently do not have a Scheduling system, so we can show system performance when only a very small amount of training data has been used. Scheduling is a task in which two subjects are given partly filled-in calendars and asked to find a common date and time to meet. It is designed to facilitate machine translation work, but is also useful for our purposes.

6.2 Evaluation Methods

We will evaluate:

- Detection of misrecognitions (and the relative contribution of each knowledge source)
- Detection / recognition of out-of-vocabulary words
- Classification of Out-Of-Vocabulary words into recognition language model classes
- Semantic classification of OOV words
- Subsequent recognition of new words
- Subsequent classification of new words

We will also measure overall recognition performance. In order to understand the effects of the new word techniques, three versions of the system will be compared

- Baseline system (no new word detection)
- New word detection with preset open classes
- New word detection with dynamic open classes (with SOUL)

The three systems will be run on a test set of spontaneous speech utterances. The performance of the systems will be reported for the entire test set as well as for the subsets consisting of utterances with new words and those without.In this way, we will be able to assess the overall effect of being able to acquire out-of-vocabulary words on system performance, specifically on accuracy and speed.

We will measure recognition word error rate and sentence error rate. The word error is the sum of the insertions, deletions and substitutions. In measuring sentence error, a sentence is wrong if any word in it is wrong. This allows us to determine how new word detection affects recognition for known words and in the presence of new words.

In addition to measuring recognition rate, we would like to measure understanding rate. Currently sites participating in the ARPA Spoken Language Systems evaluations are evaluating understanding in such systems by measuring the % correct, % incorrect and % choose-not-to-answer responses of the system. The choose-not-to-answer response means the the system rejected the utterance. In an incorrect response, the system misunderstood the utterance and did the wrong thing.

To evaluate the effects of increasing perplexity by allowing both recognition of unknown words and feedback, we will contrast the performance of all three systems under conditions where no unknown words are contained in the input.

The utility of the word learning component will be assessed by contrasting overall performance of all three systems on test sets containing unknown words. To test the added utility expected from allowing new words to be added to dynamically defined categories, as opposed to a limited set of prespecified categories, we will determine the percentage of new words learned with each of the two methods on a number of test sets which contain unknown words.

7 Bibliography

1. Adams, D.A. and Bisiani, R., ‘‘The Carnegie-Mellon University Distributed Speech Recognition System,’’ *Speech Technology,* Vol. 3, No. 2, 1986, pp. 14 - 23.

2. Asadi, A., Schwartz, R., Makhoul, J., ‘‘Automatic Modeling for Adding New Words to a Large-Vocabulary Continuous Speech Recognition System,’’ *IEEE International Conference on Acoustics, Speech and Signal Processing,* 1991, pp. 305-308.

3. Bahl, L. R., Jelinek, F., Mercer, R., ‘‘A Maximum Likelihood Approach to Continuous Speech Recognition,’’ *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. PAMI-5, No. 2, March 1983, pp. 179-190.

4. Carbonell, J. G., ‘‘POLITICS: Automated Ideological Reasoning.,’’ *Cognitive Science,* Vol. 2, No. 1, 1978, pp. 27-51.

5. Chow, Y.L., Roukos, S., ‘‘Speech Understanding Using a Unification Grammar,’’ *IEEE International Conference on Acoustics, Speech and Signal Processing,* 1989.

6. Erman, L.D. and Lesser, V.R., *The Hearsay-II Speech Understanding System: A Tutorial,* Prentice-Hall, Englewood Cliffs, NJ, 1980, pp. 340 - 360.

7. Granger, R., ‘‘A Program that Figures out Meanings of Words from Context,’’ *Proceedings of the Fifth International Joint Conference on Artificial Intelligence,* 1977.

8. Huang, X.D., Lee, K.F., Hon, H.W., Hwang, M.Y., ‘‘Improved Acoustic Modelling with the SPHINX Speech Recognition System,’’ *IEEE International Conference on Acoustics, Speech and Signal Processing,* 1991, pp. 345-348.

9. Jacobs, P., Zernik, U., ‘‘Acquiring Lexical Knowledge from Text: A Case Study,’’ *Proceedings of the Seventh National Conference on Artificial Intelligence,* 1988, pp. 739-744.

10. Kimball, O., Price, P., Roucos, S., Schwartz, R., Kubala, F., Chow, Y.-L., Haas, A., Krasner, M. and Makhoul, J., ‘‘Recognition Performance and Grammatical Constraints,’’ *Proceedings of the DARPA Speech Recognition Workshop,* 1986, pp. 53 - 59.

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

12. Lee, K.F., Hon, H.W., Reddy, R., ‘‘An Overview of the SPHINX Speech Recognition System,’’ *IEEE Transaction on Acoustics, Speech and Signal Processing,* Vol. ASSP-38, January 1990.

13. Levinson, S.E., Liberman, M.Y., Ljolje, A., Miller, L., ‘‘Speaker Independent Phonetic Transcription of Fluent Speech for Large Vocabulary Speech Recognition,’’ *IEEE International Conference on Acoustics, Speech and Signal Processing,* May 1989.

14. Levinson, S., Ljolje, A., ‘‘Continuous Speech Recognition from Phonetic Transcription,’’ *Proceedings of the DARPA Speech and Natural Language Workshop,* October 1989.

15. Lowerre, B. and Reddy, R., *The Harpy Speech Understanding System,* Prentice-Hall, Englewood Cliffs, NJ, 1980, pp. 340 - 360.

16. Nilsson, N.J., *Problem-Solving Methods for Artificial Intelligence,* McGraw-Hill, 1971.

17. Nilsson, N.J., ''Searching Problem-Solving and Game-Playing Trees for Minimal Cost Solutions,'' in *Information Processing,* A.J.H. Morrell, ed., , 1969, pp. 1556-1562.

18. Reddy, R., Newell, A., ''Knowledge and its Representation in a Speech Understanding System,'' in *Knowledge and Cognition,* L.W. Gregg, ed., L. Erlbaum Associates, 1974, pp. 256-282.

19. Stern, R.M., Ward, W.H., Hauptmann, A.G., Leon, J., ''Sentence Parsing with Weak Grammatical Constaints,'' *IEEE International Conference on Acoustics, Speech and Signal Processing,* IEEE, 1987, pp. 380-383.

20. Viterbi, A.J., ''Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm,'' *IEEE Transactions on Information Theory,* Vol. IT-13(2), 1967, pp. 260-269.

21. Ward, W.H., Young, S.R., ''Flexible Use of Semantic Constraints in Speech Recognition,'' *IEEE International Conference on Acoustics, Speech and Signal Processing,* 1993, pp. .

22. Ward, W.H., Hauptmann, A.G., Stern, R.M. and Chanak, T., ''Parsing Spoken Phrases Despite Missing Words,'' *IEEE International Conference on Acoustics, Speech and Signal Processing,* IEEE, 1988.

23. Young, S.R., ''Use of Dialogue, Pragmatics and Semantics to Enhance Speech Recognition,'' *Speech Communication,* Vol. 9, No. (5/6), 1990, pp. 551-564.

24. Young, S.R., Matessa, M., ''Using Pragmatic and Semantic Knowledge to Correct Parsing of Spoken Language Utterances,'' *Eurospeech-91,* 1991.

25. Young, S.R., Ward, W.H., ''Learning New Words from Spontaneous Speech: Automatic Detection, Categorization and Acquisition,'' *IEEE International Conference on Acoustics, Speech and Signal Processing,* 1993, pp. .

26. Young, S. R., ''Dialog Structure and Plan Recognition in Spontaneous Spoken Interaction,'' *Proceedings of the European Conference on Speech Communication and Technology,* ESCA: Paris, London, 1993.

27. Young, S. R. and Ward, W. H., ''Recognition Confidence Measures for Spontaneous Spoken Dialog,'' *Proceedings of the European Conference on Speech Communication and Technology,* ESCA: Paris, London, 1993.

28. Young, S. R. and Ward, W. H., ''Semantic and Pragmatically Based Re-Recognition of Spontaneous Speech,'' *Proceedings of the European Conference on Speech Communication and Technology,* ESCA: Paris, London, 1993.

29. Young, S. R. and Ward, W. H., ''Learning New Words from Spontaneous Spoken Speech,'' *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-93),* IEEE Press, 1993.

Table of Contents

List of Figures