Recognition Confidence Measures: Detection of Misrecognitions and Out-Of-Vocabulary Words

Sheryl R. Young

May 25, 1994
CMU-CS-94-157

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3890

An earlier draft appeared in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-94*, Adelaide, Australia, 1994. Also a version appeared in *AAAI-94 Workshop on Speech and Natural Language*, Seattle, Washington, 1994.

This research was sponsored by the National Science Foundation, under Grant No. IRI-9314992.

Keywords: machine learning, speech recognition, dialog and semantics, word learning, detecting misrecognitions and out-of-vocabulary words.

Abstract

This paper describes and evaluates a new technique for measuring confidence in word strings produced by speech recognition systems. It detects misrecognized and out-of-vocabulary words in spontaneous spoken dialogs. The system uses multiple, diverse knowledge sources including acoustics, semantics, pragmatics and discourse to determine if a word string is misrecognized. When likely misrecognitions are detected, a series of tests distinguishes out-of-vocabulary words from other error sources. The work is part of a larger effort to automatically recognize and understand new words when spoken in a spontaneous spoken dialog.

We describe a system that combines newly developed acoustic confidence measures with the semantic, pragmatic and discourse structure knowledge enbodied in the MINDS-II system. The newly developed acoustic confidence metrics output independent probabilites that a word is recognized correctly along with a measure of how reliably we can estimate if a word is wrong. The acoustic confidence metrics are derived from normalized acoustic recognition scores. The acoustic scores are normalized by estimates of the denomiator of Bayes equation. To evaluate the utility of using the acoustic techniques together with higher-level constraints, the preliminary system restricted component interaction. Words with normalized acoustic scores that had a 95% or greater probability of being incorrect were flagged prior to being input to the MINDS-II analysis module. For this study, MINDS-II *independently* used its higher-level knowledge to detect recognition errors that were semantically or contextually inappropriate. Misrecognized word strings were then re-recognized using an RTN-based speech decoder guided by a dynamically derived, highly constrained grammar that restricts the possible words that can be matched, biasing the recognizer against illogical and highly improbable content. Speaker goals and plans, contextual appropriateness, discourse and spontaneous speech structure are all considered in the derivation of grammars. A grammar is dynamically derived for each string of misrecognized words encountered within an utterance and essentially defines a set of semantic content predictions for the word string.

Although a rudimentary procedure was employed to estimate the conjoined effect of merging these two knowledge sources, the results indicate that the conjoined usage of normalized acoustic confidence measures of accuracy and the higher-level, semantic, pragmatic and discourse level constraints embodied in the MINDS-II system enables the larger system to overcomes the weaknesses of each individual technique. The techniques detect complementary phenomena. Significantly more recognition errors and out-of-vocabulary words are detected when both of these techniques are used together than when either is used alone. Alone, the acoustic methods can only detect between 2/3 and 3/4 of the recognition errors. Similarly, the higher-level constraint based methods cannot detect contextually consistent misrecogntions. Together, however, the acoustic methods detect the significant misrecognized content words that are missed by the higher-level, knowledge-based techniques. Further, the knowledge-based techniques detect most of the mid-utterance corrections and interjections as well as many of the confusible and small words that are missed by the acoustic methods. Current work focuses upon development of more sophisticated techniques for conjoining these two knowledge sources.

1. Overview

Out-of-vocabulary words constitute a major source of error in recognizing spontaneously spoken utterances. The work reported here is part of a larger project to automatically understand novel words when encountered in a limited domain, spontaneous spoken dialog. The project attempts to detect likely misrecognitions, determine if they are caused by an out-of-vocabulary word string, and if so, determine the meaning and semantic category of the new word(s) and incrementally add them to the system grammar, lexicon and semantics. Our approach to automatic out-of-vocabulary word detection and acquisition relies heavily, although not exclusively, upon being able to reliably detect misrecognized words. This has a desiriable side effect, namely that recognition errors can be found while processing input, yielding recognition confidence measures previously unavailable. Speech recognizers can output both the best matched string of words matched in an utterances and measure the probability that each is correctly matched.

Our system attempts to detect misrecognitions and out-of-vocabulary words in spontaneously spoken utterances using acoustic, semantic, pragmatic and discourse level knowledge sources. In this paper, we overview our procedures for acoustic normalization and acoustically-based, independent recognition confidence measures along with our procedures for using semantics, pragmatics and discourse analysis to detect misrecognitions. Then we describe our initial methods for merging acoustic evidence with symbolically represented semantic, pragmatic and discourse knowledge. Our methods are designed to take into account the differential reliability of each knowledge source and to arrive at an optimal decision based upon the relative power of each knowledge source in the context of the utterance. A major issue here is how to combine such knowledge sources taking into account their differential reliability under a given set of conditions. We extend the Bayesian techniques we used for estimating acoustic confidence to other knowledge sources as well. This allows us to optimally weight each knowledge source. We assign confidence measures to words (and phrases) that take into account the power of each knowledge source under the specific set of conditions.

This work extends our previous work on detecting misrecognized and out-of-vocabulary words by normalizing word scores with an all-phone representation of the input [1, 2]. It also extends our work on using discourse [3], semantics and pragmatic knowledge to detect misrecognitions and to generate predictions used in re-recognizing select substrings [4, 5]. It also extends the work of Asadi et.al. [6] in detecting out-of-vocabulary words. The work illustrates how differential knowledge source reliability can be assessed and how knowledge sources can be merged to detect out-of-vocabulary words.

2. Acoustic Confidence Metrics

Current Hidden Markov Model (HMM) based speech recognizers use an evaluation function based on Bayes' equation to score hypotheses. They produce the most probable word sequence given the acoustic input according to the formula:

$$P(W|A)= P(A|W)\, P(W) / P(A).$$

Here P(A|W) represents the probability of the acoustic sequence given a word string. This probability is provided by the HMMs for the words in the lexicon. P(W) represents the apriori probability of the word string and is usually provided by a stochastic language model (bigrams or trigrams). P(A) represents the apriori probability of the acoustic sequence.

Most speech recognition systems do not attempt to estimate P(A). The rationale for ignoring the apriori probabilities of acoustic sequences is that P(A) is the same for all utterances in a time synchronous decoding. Hence, the estimation of P(A) will not change the relative order of word string hypotheses output by a speech recognizer, and therefore will not change which one is picked as best. This means that the scores assigned by speech recognizers to word and sentence hypotheses are not absolute measures of probability, but rather relative measures. We know which utterance is most likely, but don't really know how good of a match it is. In other words, we have no true measure of goodness of match and have no real means for evaluating accuracy of output word strings.

However, it is necessary to be able to evaluate confidence in recognition to be able to detect new words and to know when to engage in a clarification dialog with a user. Although it is possible to generate a "generic" novel word model [6] and design the language model so that the generic word model competes with other, known word hypotheses, such models do not provide a measure of goodness of match and cannot be combined with other knowledge sources to optimize match. There are many possible sources of information in a speech understanding system to help estimate confidence in a hypothesis, including semantics, pragmatics, discourse structure, acoustic ambiguity, syntax, structure of spontaneous speech, etc. Each of these knowledge sources can reliably detect certain types of information, yet each has its relative weaknesses.

In order to combine the information from various sources in an optimal manner, we must be able to estimate the reliability of each piece of information in the current context. In other words, we can develop differential reliability and differential error models for each knowledge source used in a spoken language system. Such models indicate, for each potential, modeled knowledge source, the types of phenomena most reliably detected and the characteristics of phenomena where the knowledge source is unreliable.

2.1. Word Score Normalization

To assess how well the system is able to reject misrecognitions using acoustic information alone, we developed a technique for acoustic normalization and then evaluated it on data to assess the differential reliability and detection power of our acoustic knowledge source. We sought to characterize acoustic phenomena in terms of both how well we can reject incorrect acoustic hypotheses and how reliably we can match the phenomena.

Most speech recognizers produce a maximum likelihood word sequence using acoustic models and word-level language models. They output either the single best hypothesized word string or the n-best word strings. Normally, only these word strings are considered when performing later processing such as inferring utterance meaning. The scores assigned by the recognizer are a weighted sum of the log probabilities from the acoustic and language models. They work by maximizing the most likely word string path. Paths are extended by computing acoustic match scores for each potential word that can extend a path and merging this information with the prior path score and the language model transition probability for the individual word. Those path extensions that result in the best overall score are retained for further extension, while those falling below a certain threshold are pruned and not considered further. The scores produced are not normalized. As discussed above, they do not represent any absolute measure of the match, but are meaningful only in comparison to other hypotheses produced for the same utterance. The score produced by the recognizer is therefore not really useful directly for rejecting utterances or regions of utterances as misrecognitions. It can only be used for selecting among utterance hypotheses and can only be used to compose hypotheses from known or directly modeled words.

We developed a method that enables us to directly assess the confidence of an acoustic match. To do this, we begin by normalizing the scores output by the recognizer, transforming the scores so that they take into account overall goodness of recognition. We use a phone-based decoding as a basis for normalizing the word-based decoding.

To normalize the word score produced by the recognizer, we subtract the log-probability score for an all-phone recognition from the log-probability word score and normalize for length. The all-phone score is generated by running the speech recognizer on the utterance allowing any triphone to follow any other triphone with a trigram probability for triphone sequences. A triphone is a context dependent phone model. Trigrams of the triphone sequences are computed from a large corpus of English language text. We use Bayesian Updating to turn the normalized word score into a confidence measure. For this, words can be grouped into classes or estimated individually. For each word (class) we estimate when a word is seen with a particular score, what is the percentage of time that the word was correctly recognized. This estimate is made by running the recognition system on a training set of data. This gives us a direct measure of the confidence with which we can reject or accept a word based on acoustic measures.

A phone-based decoding search is run in parallel with the word-based search. The phone decoding uses bigrams of phone transitions as a language model in the same way that the word search uses bigrams of word transitions. In order to normalize a word score, the score from the phone path for the same set of frames is subtracted from the word score. Since the scores are log-probabilities, this subtraction represents a division of probabilities. The result of the subtraction is then divided by the length of the word in frames (10 msec increments). The acoustic match scores in the word search are constrained by word sequences from the language model and phone sequences from word models. The phone search provides an estimate of the acoustic match of phone models to the input unconstrained by word or word-sequence models. The phone search is constrained only by phone sequences characteristic of the language (English) without respect to the current lexicon or language model.

2.2. Experiment 1

In order to determine if the normalization procedure provides a more useful score that the relative sources normally output by a recognizer, we performed an experiment using spontaneous speech from the ATIS training corpus. 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 the SPHINX-I discrete HMM-based speech recognizer [7] with a word bigram language model. SPHINX-I outputs a single best word string for each recognized utterance. The test utterances were spontaneous spoken speech, and included noise such as filled pauses, (uhmms, ahms), stutters and partial words, as well as ill-formed utterances, mid-utterance corrections and restarts. Our system directly models noise (filled pauses, stutters, partial words) [8, 9] and uses a semantically-based phrase recognition algorithm for processing ill-formed and edited utterances [10].

To assess ability to correctly reject misrecognized words, we followed the following procedure. For the words in the hypotheses output by the recognizer, we saved the acoustic word scores and flags indicating whether the words were correct. Correctness was determined by aligning the words in the hypotheses with transcripts for the utterances. From this data, we created signal (correct) and noise (incorrect) distributions for each word. From these we estimated our ability to reject words by looking at the overlap of the signal and noise distributions for each of the 1800 words in the lexicon. We assessed the system's ability to correctly reject misrecognitions looking at the measures of d-prime, D', and power. D' measures the difference between the means of the signal and noise distributions. The larger the D', 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.

The average power for the 1800 words in the lexicon using regular acoustic scores was 65%. We then normalized the word scores according to the above procedure and calculated the average power. For the normalized scores, the average power increased from 65% to 74%. The results for the normalized scores are depicted in Table 4-2. The results indicate that, in general, the normalization procedure makes correct and incorrect words more separable.

However, we need to note that word recognition rates and power both vary widely across words. Some words can be reliably discriminated. However, there are words that can be correctly recognized most of the time but not reliably rejected because misrecognitions occur (infrequently) throughout the range of normalized scores. Further there are words whose baseline recognition rates are low but where we can correctly reject misrecognitions almost all the time and words which are neither correctly recognized or rejected, as illustrated in Table 2-1. By and large, longer words and unique words are well discriminated while very short, non-distinct words and function works cannot be reliably rejected when they are incorrectly recognized. We found that this normalization was a good discriminator for some words but not for others and in general still doesn't provide a good confidence measure. The correct and incorrect distributions for some words were very distinct, while for others were highly overlapped. Also, this measure

Correct Acceptance and Correct Rejection Rates for Words							
Words	C. Accept	I. Reject	Freq.	Words	C. Accept	I. Reject	Freq.
san_francisco	1.00	1.00	282	hi	1.00	0.91	22
pittsburgh	1.00	1.00	110	will	0.87	0.50	54
airport	1.00	1.00	98	me	1.00	0.73	34
information	1.00	1.00	68	the	1.00	0.00	916
american	1.00	1.00	60	on	1.00	0.00	640
saturday	1.00	1.00	26	a	1.00	0.00	284
when	1.00	1.00	22	have	0.96	0.31	132
price	1.00	1.00	20	what	0.98	0.29	308
sixty	1.00	1.00	18	is	0.96	0.26	236
delta	0.98	1.00	98	and	1.00	0.11	358
airlines	0.97	1.00	171	how	0.98	0.11	138
i'm	1.00	0.95	188	to	1.00	0.00	1114

Table 2-1: Accurate Acceptance and Rejection Performance Rates

doesn't account for the frequency of correct vs incorrect words with a given score, it only uses the percentage of the area under each of the two curves. So, while these scores may be useful for rejection, they still don't provide a direct measure of confidence in the correctness of the word given its score. In order to turn the normalized score into a confidence measure we use a Bayesian updating method to estimate the probability that a word is correct when it has a given score.

2.3. Experiment 2: Acoustic Probabilities

We estimated the acoustic probability that a word is correct with a given normalized score for each of the 1800 words in the lexicon, in spite of the fact that we did not have enough data to make such estimates reliably for every word. None-the-less, we conducted the following experiment without clustering words, relying exclusively on the signal (correct recognition) and noise (incorrect) disctributions computed above.

To compute probabilities, for each word, we quantized the range of normalized scores into 75 bins or score ranges. We then took normalized word scores from 5000 utterance recognition hypotheses (~30,000 words) taken from the ARPA ATIS2 training data described above, and accumulated histograms for each word. For each bin associated with a word, we determined the percentage of the time word was correct when its normalized score was in the bin. 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.

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. As described previously, the Sphinx-I discrete HMM speech recognizer was to generate the word hypotheses using a lexicon of approximately 1800 words, including ten non-verbal events, and used a word-class bigram with a perplexity of 55. We set a rejection criteria to maintain 95% correct accepts and determined the ability to reject misrecognitions. As a test set, we used the ARPA Feb92 ATIS test set, containing 1000 utterances from speakers not seen in training.

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. These results are shown in Table 4-1 In looking at the histograms for the word classes, some had almost perfect classification, while others had only slightly better than chance. So for some word classes, we can very reliably accept correct words and reject misrecognitions on acoustic evidence.

In summary, the evidence suggests that the acoustic normalization technique and the acoustic confidence measures can reliably reject a significant number of misrecognized words. The set of words that are reliably rejected tend to be semantically unique, as opposed to function words or very short, high frequency words. Next, we wish to capitalize upon this ability to reliably reject some misrecognized words and see if we can use it to augment the capabilities of the semantic, pragmatic and discourse-level constraints on the recognition process. Specifically, we hope that the discourse-based module will detect those recognition errors missed by the acoustic module and that the acoustic module will be able to detect phenomena to which the semantic-pragmatic-discourse module is insensitive .

3. Semantic, Pragmatic and Discourse Based Discrimination

The various MINDS systems use higher-level, knowledge-based techniques to constrain recognition. The systems operate by analyzing input and dynamically generating constraints that define content that is reasonable, meaningful or logical given prior discourse. [11] These constraints are the translated into system grammars that are used to guide the normal speech decoding process, in a manner similar to a standard language model. In other words, the system applies meaning and structure based constraints to restrict the possible words that can be matched during the decoding process. The MINDS-II system [5] operates using the following loop:

- Spontaneously spoken input is digitized and recognized using a standard statistical language model and an HMM-based recognizer.
- The recognized string is semantically parsed. [12]
- MINDS-II evaluates both the recognized string and its semantic parse.
    - It corrects inaccurate or incomplete semantic representations.
    - It detects inappropriate content or likely misrecognitions that violate contextual constraints.

- Content predictions are generated for each misrecognized word string within an utterance.
- Content predictions are translated into semantically-based RTN recognition grammars.
- Each misrecognized word string (and competing start-end sequences of word strings) within an utterance is re-recognized using an RTN-based HMM decoder and the appropriate recognition grammar. [13]

The MINDS-II system analyzes all input and looks for both parse errors and likely misrecognitions. Its analysis is based upon semantics, pragmatics and discourse structure constraints. Specifically, it first looks to see if the words within an utterance make sense relative to one another. Here, the structure of spontaneous speech is considered. For example, the system has a set of heuristics for recognizing restarted utterances and mid-utterance corrections. The system also considers the meaningfulness of the utterance in terms of prior context and information introduced by either the speaker or their partner or database backend. The system looks to see whether a speaker references information that is available (vs. unavailable) for reference. It also evaluates how the utterance furthers the speakers goals and plans and determines the type of discourse plan embodied in the utterance. The system has a set of heuristics and algorithms for traversing both domain plans and discourse plans. [2] These heuristics constrain both the types of discourse plans available at each point in the dialog and the content of these respective discourse plans. Should the system find any information that violates any of the above heuristics, it attempts to identify which words are most likely to be responsible for the violation using abductive reasoning. When one or more strings of words within an utterance are flagged, the system works to define the set of possible semantic contents that make sense given all of the context and the structure of the discourse and of spontaneous spoken utterances. This process is responsible for generating "predictions" that are used to constrain the re-recognition process.

Predictions are generated by defining all possible semantic content that could have been said and still make sense. In contrast to abductive reasoning, the system does not attempt to define the best. Rather, its goal is to be inclusive, to define the complete set of what is possible given each applicable discourse and domain plan step. Usually, the initial analysis of the input utterance will result in the identification of a single discourse (and if appropriate, a single domain plan) step, although the system uses multiple, competing discourse and domain actions to compute the set of possible content when necessary. The set of concepts included in the final predictions satisfy all constraints for each possible "condition". For example, if a mid-utterance correction could have occurred in words 4-6, the system will compute all concepts available for modification from words 1-3 and contained in the last (embedded) constituent given the prevailing set of discourse and domain plans and the constraints on what information is available for reference. If the prior constituent contains an embedded concept, either the entire consitutent or just the last embedded concept could be modified or refined in the mid-utterance correction.

Predictions regarding content are translated into a grammar that is used to guide an RTN-based decoder that re-processes the misrecognized words identified earlier. The grammars are highly constrained and restrict the possible words that can be matched during the decoding process. In other words, the recognizer is biased against illogical and highly improbable content. There is a set of predictions generated for each word string within an utterance that could be misrecognized. The predictions are dynamically generated and designed to apply all applicable constraints upon semantic content.

3.1. Strengths and Weaknesses of Semantic Module

In order to determine how to use the acoustic evidence in conjunction with our existing semantic, pragmatic and discourse based analysis system (MINDS, MINDS-II) we needed to evaluate the relative strenghts and weaknesses of the knowledge-based module. To do this, we evaluated the ability of the MINDS-II system to detect recognition errors using the same recognizer, lexicon, phone models, word-bigrams and test set used in the acoustic experiments. In addition, we evaluated performance on two additional ARPA ATIS test sets that contained 1,000 spontaneous spoken utterances apiece.

Again, the MINDS-II semantic-pragmatic-discourse analysis system used a lexicon of approximately 1800 words, including ten non-verbal events, a word-class bigram were trained on approximately 12000 utterances taken from the DARPA ATIS2 training set, and have a perplexity of approximately 55. There are 79 concept nets. The results show the analysis system’s ability to detect recognition errors and its ability to correctly predict the content or meaning of the misrecognized word strings.

The system processed all of the dialogs in each of three ARPA ATIS test sets. The parses from each spoken utterance were passed to our ATIS back end, which parsed the string and produced a response from the database. The ARPA test sets randomly assess performance on a subset of the input utterances, ensuring that "Class X" queries, or those queries for which there is no reference answer, are not included.

Table 3-1 shows the performance results of the ARPA ATIS test set used in the acoustc experiments> The results are representative of performance on all three (3,000 utterances) test sets. As shown in Table 3-1, overall error rate can be

Error Type	Initial Error	Detected Errors	Correct Predictions	Remaining Error
Inconsistent Context	11.81	10.46	10.01	1.8
All Errors	20.58	10.46	10.01	10.57

Table 3-1: Semantic, Pragmatic and Discourse Based Error Detection and Correction

divided into contextually consistent and contextually inappropriate word recognition errors. The MINDS & MINDS-II systems (semantic/pragmatic/discourse module) cannot detect contextually consistent word substitutions. Contextual appropriateness is defined in terms of the discourse plans that can be executed at a specific point in time (e.g. clarify, confirm, correct, continue), the objects, attributes and actions available for reference given prior dialog context, and the goals and plan steps that are active or currently under discussion. For example, if a flight number is misrecognized and substituted for another flight number that fulfills the same semantic constraints previously specified in the dialog (i.e. both go to the same place / leave at the same time / serve a meal, etc.) it is considered to be a semantically consistent recognition error and cannot be detected by the semantic, pragmatic and discourse knowledge available in the current two-pass recognition system. Overall, in the three ARPA test sets evaluated (approximately 3,000 utterances) roughly 39% of all errors are not detectable by the semantic/discourse module. With the Sphinx-I recognition system, this corresponds to roughly a 9% error rate.* These contextually consistent errors can only be detected using other knowledge sources.

The majority of errors are contextually inappropriate, and it is on these errors that we can measure the sensitivity of the semantic, pragmatic and discourse knowledge and evaluate the strengths and weaknesses of the approach. As seen in Table 3-1, the system can both detect (*detected errors*) and generate accurate predictions (*correct predictions*) for most of the semantically inconsistent errors. Specifically, the system generated correct predictions for 88% of the contextually inconsistent errors, correctly predicting semantic content and translating the predictions into a recognition lexicon and grammar.

4. Combining Knowledge Sources: Interactions

The acoustic normalization and acoustic-based confidence measures as well as the semantic-pragmatic-discourse based analysis methods can each detect recognition errors, regardless of their underlying cause. Each method has relative strengths and weaknesses. In this section we discuss how to capitalize upon the relative strengths of the two methods and combine them to enhance overall system performance.

Acoustic Decision	**Correct Recognition**	**Incorrect Recognition**
Accept	.94	.46
Reject	.06	.54

Table 4-1: D’ Results using Acoustic Confidence Measures on ARPA Test Set N=1000 utterances, Experiment 2

Acoustic methods can detect misrecognized words [2, 1]. A Bayesian Updating

*Different HMM based recognition systems have different overall error rates. However, we expect that the percentage of errors that are semantically consistent to be roughly equal when using any of todays state-of-the-art recognizers.

paradigm is used to estimate recognition confidence based on normalized acoustic word scores. As seen in Table 4-2, for some words, namely 74%, correct hypotheses can be reliably discriminated from incorrect ones using normalized acoustic scores. For the remaining 26% of words, such a discrimination cannot be made reliably. The acoustic confidence metric assesses when reliable decisions can be made and when they cannot be made. For those words where misrecognitions can't be discriminated on acoustic evidence, some other form of evidence must be used. Previous work shows that semantic, pragmatic and discourse level constraints can detect many misrecognitions [4, 5]. However, in these results, roughly 39% of the recognition errors were consistent with all semantic, pragmatic and dialog constraints.

Given the goal of this study was to evaluate whether the conjoined use of acoustic confidence measures together with semantic-proagmatic-discourse based methods would significantly improve a system's ability to correctly reject misrecognized input, we evaluated a test set consisting of 70 dialogs taken from an ARPA ATIS test set. The evaluated dialogs contained 4,319 words.

To merge the acoustic and knowledge-based analysis techniques for this preliminary evaluation, we added the acoustic confidence metric into our existing dialog system, MINDS-II, which detects misrecognitions and attempts to re-recognize misrecognized input using a dynamically derived, limited lexicon and grammar. Specifically, the MINDS-II system was input with the words flagged by the acoustic confidence module. MINDS-II then performed its normal processing of the input, flagging those recognition errors it detected, generating appropriate predictions for the misrecognized regions and then performing a prediction-based re-recognition of the flagged, misrecognized substrings. Finally, the misrecognitions detected by both modules were summed. The MINDS-II system did not receive the confidence measures associated with each of the recognized words and use these in its processing of the input. Although the more sophisticated method of reasoning using the acoustic confidence scores is advisable and should enhance performance, the goal of this study was to see whether the two knowledge sources would detect complementary sets of misrecognitions.

Specifically, we wished to determine whether each of the modules would detect misrecognized input missed by the other. To determine how to combine the knowledge sources, we examine the relative strengths and weaknesses, or the types of errors associated with each approach. We know that the MINDS / MINDS-II system can accurately detect semantically or contextually inappropriate misrecognitions using dialog-based and semantic knowledge-based techniques. More importantly, these techniques do not inaccurately flag input. However, they also do not detect a significant percent of misrecognitions. Hence, we evaluated whether the acoustic methods can detect misrecognized input that is semantically and contextually consistent. Further, we evaluated the areas where both the semantic/pragmatic/discourse component and the acoustic measures both indicated that the recognized input is erroneous.

The acoustic techniques are limited by the number of words that they can reliable dscriminate (74%) and their tendency to incorrectly reject correctly recognized words. The results from Experiments 1 and 2 indicate that the misrecognized words that are not reliable detected with the acoustic methods are high frequency, short words. The semantic and pragmatic techniques detect sequences of words that do not make any sense and word sequences that form interjections, mid-utterance corrections and restarts. Apriori, it appreared that the acoustic techniques would be likely to detect the semantically consistent misrecognitions missed by the semantic and pragmatic techniques. The acoustic error patterns are illustrated in Tables 4-1 and 4-2. The same analysis is presented for the

Acoustic Accept	**Correct Recognition**	**Incorrect Recognition**
Accept	.95	.26
Reject	.05	.74

Table 4-2: D' Results for Acoustic Normalization for 1800 Words, Experiment 1

semantic/pragmatic/discourse knowledge source in Table 4-3.

Semantic Decision	**Correct Recognition**	**Incorrect Recognition**
Accept	1.0	.41
Reject	.00	.59

Table 4-3: D' Results for Semantic/Pragmatic/Discourse Component

What is important to note here is that we want the semantic/pragmatic/discourse component in the merged system to decrease the false acceptance rate associated with the acoustic knowledge without increasing the rate at which correctly recognized input is rejected. Since the semantic/pragmatic componenent does not falsely reject accurate input, it should not. Similarly, we want the acoustic component to capture misrecognized input that is contextually consistent and is not detectable by the knowledge-based analysis component and the knowledge-based component to detect misrecognitions in the words that cannot be reliably rejected.

4.1. Decision Rules for Combining Knowledge Sources

Given these sets of results and the error patterns associated with each of the modules, we decided to begin experimenting with the following decision rule for combining the two knowledge sources. As illustrated in Table 4-4, our rule has two parts. First, if the semantic/discourse module rejects a word string (or phrase) and decides that it is misrecognized, we will reject the string regardless of its probability correct. Second, if the acoustic module rejects a word string we will reject it even if the semantic module says to accept it. In this way, we will not increase nor decrease the false rejection rate (rejecting correctly recognized words) but we do stand to significantly decrease the inaccurate acceptance rate.

Acoustic Decision	Semantic Decision Accept	Semantic Decision Reject
Accept	**ACCEPT x 2**	**REJECT**
Reject	**REJECT**	**REJECT x 2**

Table 4-4: Decision Rules for Combining Acoustic and Semantic/Discourse Decisions on Accuracy of Recognized Word String

4.2. Results of Combined Knowledge Sources

We evaluated how well the acoustic normalization technique complemented the semantic/pragmatic/discourse based methods and their conjoined effectiveness at detecting misrecognized words. The test set was composed of 70 dialogs randomly taken from the ARPA ATIS Oct. 92 test set. The test set contained 4,319 words including a number of words that were unknown to our system. All out-of-vocabulary words had never been seen or represented in the system. We used a modified version of the SPHINX-I recognizer to recognize, acoustically normalize and re-recognize input using a prediction-based, constrained grammar. Of course, all out-of-vocabulary words resulted in word substitution errors were sequences of known words are substituted for the correct (unknown) word strings. In these cases, the system's objective is to detect the misrecognized word substitutions.

Error Rate Anaysis			
Initial Error	Following Acoustics	Following Semantics	Following Both
11.0	8.3	3.9	2.7

Table 4-5: Error Rate Reductions Following Misrecognition Detection, 70 dialogs

The results of this preliminary analysis show that these two knowledge sources can and do detect a complementary set of misrecognized input. Specifically, on this set of 70 dialogs, composed of 4,319 words that were recognized by the SPHNIX-I system, 73.1% of all misrecognitions or all but 2.7% of the errors were detected by the conjoined use of the acoustic normalization / confidence measures and the MINDS-II semantic-pragmatic-discourse module. This is a significant improvement in performance relative to each of the individual systems (acoustic normalization alone and MINDS-II semantic/pragmatic/discourse analysis and re-recognition alone) abilities to detect misrecognized input. The conbined system detected 19% and 14% more misrecognitions than the acoustic and semantically based systems respectively.

The two error detection methods did not tend to detect the same errors. Specifically, only 15% of the misrecognized words were flagged by both modules. The acoustic module detected semantically significant content words that were misrecognized. Many of these were sematically consistent with the surrounding utterance and discourse context and therefore missed by the MINDS-II module. The acoustic module was not able to detect many of the small words that made composed interjections, mid-utterance corrections and restarts. As the MINDS-II system was designed with knowledge of spontaneous speech patterns and is equipped with algorithms to detect mid-utterance corrections and restarts, those phenomena were readily flagged. Similarly, the MINDS-II module is designed to evaluate the semantic content of a recognized utterance with respect to the preceeding and surrounding semantic constraints, discourse plans, domain plans, speaker goals and discourse structural constraints, it can identify misrecognied input that is inconsistent. Both the MINDS-II system and the acoustic module had difficulty with confusing contractions with their associated expansions as well as with injections and insertions of small words such as ***"a" "the" "me" "in" "do" "will" "and" "on" "all"***.

Joint Decision	Correct Recognition	Incorrect Recognition
Accept	.95	.27
Reject	.05	.73

Table 4-6: D’ Results for Combined Knowledge Sources, 70 dialogs

Given these results, and the success of this initial investigation, we now plan to go ahead and more thoroughly integrate the acoustic confidence module with the MINDS-II system. To begin, we intend to feed the exact accoustic correct probabilities associated with each recognized word into the MINDS-II system for analysis. This gives the MINDS-II system more information to reason upon, may enhance its ability to detect misrecognized words and may enable the MINDS-II system to look more closely at those words the acoustic module scores as not-very-probably correctly recognized, perhaps decreasing overall false rejection scores. Second, we plan to investigate more sophisticated methods for merging these two knowledge sources. The initial decision rule employed in this study can be improved upon. We hope that future research will enable us to both maintain the advantages of combining these two knowledge sources while providing us with ways to decrease the false rejection rate of the acoustic module, without decreasing the ability to acoustically identify misrecognized words. Third, we intend to investigate the use of normalized acoustic scores during the highly constrained, re-recognition process. Given that the grammars used to guide decoding during re-recognition of strings within an utterance are low in perplexity, it seems likely that the acoustic normalization process will better discriminate incorrectly recognized words. Finally, we are investigating methods for directly modelling out-of-vocabulary words in the language model and the use of a modified bigram-semantic grammar to guide the re-recognition process.

REFERENCES

1. 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.

2. 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.

3. 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.

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

5. 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.

6. 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.

7. 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.

8. Ward, W.H., ‘‘Modelling Non-Verbal Sounds for Speech Recognition’’, *Proceedings of the DARPA Speech and Natural Language Workshop*, October 1989.

9. Wilpon, J., Rabiner, L., Lee, C.-H., Goldman, E., ‘‘Automatic Recognition of Keywords in Unconstrained Speech using Hidden Markov Models’’, *IEEE Transaction on Acoustics, Speech and Signal Processing*, Vol. ASSP-38, No. 11, 1990, pp. 1870-1878.

10. Ward, W., Issar, S., Huang, X., Hon, H., Hwang, M., Young, S. R., Matessa, M., Stern, R. and Liu, F., ‘‘Speech Understanding in Open Tasks’’, *Proceedings of the DARPA Speech and Natural Language Workshop*, Morgan Kaufmann, 1992.

11. Young, S.R., Hauptmann, A.G., Ward, W.H., Smith, E.T., Werner, P., ‘‘High Level Knowledge Sources in Usable Speech Recognition Systems’’, *Communications of the ACM*, Vol. 32, No. 2, 1989, pp. 183-194.

12. Ward, W., ‘‘Understanding Spontaneous Speech: The PHOENIX System’’, *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1991, pp. .

13. Ward, W. H. and Young, S. R., ‘‘Flexible Use of Semantic Constraints in Speech Recognition’’, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-93)*, IEEE Press, 1993.

Table of Contents

1. Overview	**3**
2. Acoustic Confidence Metrics	**3**
2.1. Word Score Normalization	**4**
2.2. Experiment 1	**6**
2.3. Experiment 2: Acoustic Probabilities	**7**
3. Semantic, Pragmatic and Discourse Based Discrimination	**8**
3.1. Strengths and Weaknesses of Semantic Module	**10**
4. Combining Knowledge Sources: Interactions	**11**
4.1. Decision Rules for Combining Knowledge Sources	**13**
4.2. Results of Combined Knowledge Sources	**14**
REFERENCES	**16**

List of Tables

Table 2-1:	**Accurate Acceptance and Rejection Performance Rates**	**7**
Table 3-1:	**Semantic, Pragmatic and Discourse Based Error Detection and Correction**	**10**
Table 4-1:	**D' Results using Acoustic Confidence Measures on ARPA Test Set N=1000 utterances, Experiment 2**	**11**
Table 4-2:	**D' Results for Acoustic Normalization for 1800 Words, Experiment 1**	**13**
Table 4-3:	**D' Results for Semantic/Pragmatic/Discourse Component**	**13**
Table 4-4:	**Decision Rules for Combining Acoustic and Semantic/Discourse Decisions on Accuracy of Recognized Word String**	**14**
Table 4-5:	**Error Rate Reductions Following Misrecognition Detection, 70 dialogs**	**14**
Table 4-6:	**D' Results for Combined Knowledge Sources, 70 dialogs**	**15**