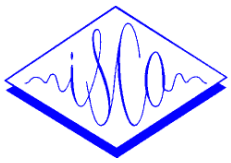

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

ESTIMATING RECOGNITION CONFIDENCE: METHODS FOR CONJOINING ACOUSTICS, SEMANTICS, PRAGMATICS AND DISCOURSE **

Sheryl R. Young
School of Computer Science, Carnegie Mellon University
Pittsburgh, PA 15213-3891 USA

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 utterances and dialogs using multiple stochastic and symbolic knowledge sources including acoustics, semantics, pragmatics and discourse structure. The work is part of a larger effort to automatically recognize and understand new words when spoken.

The system described combines newly developed acoustic confidence measures with the semantic, pragmatic and discourse structure knowledge embodied in the MINDS-II system. The acoustic confidence metrics output independent probabilites that a word is recognized correctly and measure how reliably we can estimate if a word is wrong. The acoustic confidence metrics are derived from normalized acoustic recognition scores. 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 and dynamically derived language model that biases against recognition of illogical and highly improbable content. The dynamically derived grammars restrict the words that can be matched during recognition, reducing perplexity by defining a set of semantic content predictions for the word string. A grammar is derived for each misrecognized word string encountered within an utterance. Speaker goals and plans, contextual appropriateness and structural characteristics of discourse and spontaneous speech are all considered in the derivation of grammars.

The results indicate that the conjoined usage of acoustic confidence measures of accuracy and higher-level constraints increased ability to detect misrecognitions by 36% and enabled the larger system to overcomes the weaknesses of the individual techniques. The techniques detect complementary phenomena. The acoustic methods detect important, misrecognized content words. They cannot reliably estimate recognition accuracy for most small or confusible words. The higher-level constraint methods cannot detect contextually consistent misrecogntions, but can detect errors caused by confusible content words, restarts and mid-utterance corrections. Current work focuses upon development of more sophisticated techniques for conjoining these two methods and techniques to use acoustic confidence measures during decoding.

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 a novel 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. Systems can use confidence measures to guide further system processing and interaction with speakers. Further, recognition confidence measures can be used during decoding to potentially improve recognition performance.

This work extends our previous studies on detecting misrecognized words acoustically [1, 2]. In that work we used a Bayesian Updating paradigm to estimate recognition confidence based on normalized acoustic word scores. The study showed that for some words, correct hypotheses could be reliably discriminated from incorrect ones using normalized acoustic scores. For other words, such a discrimination could not be made, but the confidence metric enables a system to know when it can reliably make a decision. For those words which can't be discriminated using acoustic evidence, some other form of evidence must be used. Previous work shows that semantic, pragmatic and discourse level constraints can detect many misrecognitions [3]. However, higher-level knowledge sources cannot discriminate misrecognized words that are consistent with all constraints. This paper reports on a system which attempts to detect misrecognitions and out-of-vocabulary words using acoustic, semantic, pragmatic and discourse level knowledge sources together. It reviews our procedures for performing acoustic normalization and deriving acoustic confidence measures as well as our procedures for using higher-level constraints to detect and correct misrecognized input. Then it describes our initial methods for merging stochastic acoustic evidence with symbolically represented higher-level knowledge and resulting system's performance. 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 is how to combine such knowledge sources, assess their reliability and use them to assign confidence measures to words and phrases.

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

2. 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 [4] 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 potential, modeled knowledge source used in a spoken language system. Such models indicate the types of phenomena most reliably detected and the characteristics of phenomena where the knowledge source is unreliable.

3. ACOUSTIC 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 its differential reliability and detection power.

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. 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 words that are misrecognized. 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. The technique is based upon normalizing the scores output by the recognizer, transforming the scores so they take into account overall goodness of recognition. We used 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 (context dependent phone model) to follow any other triphone with a trigram probability for triphone sequences. 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.

3.1. Word Score Normalization Experiment

To evaluate whether 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 [5] 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) [6, 7] and uses a semantically-based phrase recognition algorithm for processing ill-formed and edited utterances [8].

To assess ability to correctly reject misrecognized words, we used 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 word string hypotheses with transcripts for the utterances. From this data, we created signal (correct) and noise (incorrect) distributions for each word. To estimate the systems ability to reject words we looked at the overlap of the signal and noise distributions for each of the 1800 words in the lexicon. To measure the system's ability to correctly reject misrecognitions we used 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.

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

Table 1: D' Results for Acoustic Normalization of 1800 Words. EXP. 1

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 1. The results indicate that, in general, the normalization procedure makes correct and incorrect words more separable.

However, both word recognition rates and power vary widely and independently across words. Some poorly recognized words can be reliably rejected while others cannot. 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 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 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. 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.

3.2. 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, this experiment was conducted 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 to yield 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 and, in contrast to Experiment 1, reflect word frequency effects in the test set. Again, the Sphinx-I discrete HMM speech recognizer was used to generate the word hypotheses using, roughly, an 1800 word lexicon, including ten non-verbal events, and a word-class bigram of perplexity 55. We set a rejection criteria to maintain 95% correct accepts and determined the ability to reject misrecognitions. The test set used was the ARPA Feb92 ATIS test set, containing 1000 utterances from speakers not seen in training.

Results for this test set are shown in Table 3. 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 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 reliably discriminated tend to be semantically unique, as opposed to function words or very short, high frequency words. We hope to capitalize upon this ability to reliably reject many misrecognized words acoustically and evaluate whether we can augment the error detection capabilities of our higher-level knowledge-based system. 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.

4. HIGHER-LEVEL 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. [9] These constraints are 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 [3] operates in 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. [10]
- MINDS-II evaluates 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 words) is re-recognized using an RTN-based HMM decoder [11] and an RTN recognition grammar.

The MINDS-II system analyzes all input and looks for both parse errors and likely misrecognitions using semantics, pragmatics and discourse structure constraints. Initially, it looks to see if the words within an utterance make sense relative to one another. Here, the structure of spontaneous speech is considered using a set of heuristics for recognizing restarted utterances and mid-utterance corrections. Next, the system considers the meaningfulness of the utterance in terms of prior context and information previously introduced 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. [1] 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 structural discourse constraints. 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 content when applicable. 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. A set of predictions is dynamically generated for each word string within an utterance that could be misrecognized.

4.1. 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, a lexicon of approximately 1800 words, including ten nonverbal events, was used as was the word-class bigram of perplexity 55 trained from roughly 12000 DARPA ATIS2 training utterances. The system used 79 RTN concept nets for semantic parsing (PHOENIX) and for translating content predictions into a grammar/language model. 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. Performance was evaluated using the dialogs in each of three ARPA ATIS test sets. The complete system, including recognizer, semantic parser, MINDS-II analysis, RTN recognizer and the ATIS back end, was used in the evaluation.

Table 2 shows the performance results of the ARPA ATIS test set used in the acoustic experiments. The results are representative of performance on all three (3,000 utterances) test sets. As shown in Table 2, overall error rate can be divided into contextually consistent and contextually inappropriate word recognition errors. 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. The MINDS systems 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. Using today's state-of-the-art HMM-based recognizers, 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.

Error Analysis: Misrecognitions Detected				
Error Type	Initial Error	Detected Errors	Correct Predict	Final Error
Inconsistent Context	11.81	10.46	10.01	1.8
All	20.58	10.46	10.01	10.57

Table 2: Higher-level Constraints: Error Detection and Correction

* These contextually consistent errors can only be detected using other knowledge sources.

5. CONJOINING KNOWLEDGE SOURCES

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. How can we combine then to capitalize upon their relative strengths and maximize overall system performance?

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

Table 3: D' Results for Acoustic Confidence Measures on ARPA Test Set, Exp. 2

Normalizing acoustic scores enables detection of misrecognized words and provides the raw input for estimating recognition confidence using a Bayesian Updating paradigm. As seen in Table 1, for some words, namely 74% of those in our lexicon, 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. In this study, we evaluate whether higher-level constraints can be used to detect a complementary set of misrecognized words. Given these objectives, we evaluated the conjoined system described below using the same ARPA ATIS test set used in the earlier, isolated component evaluations.

A preliminary version of a conjoined system that merges the acoustic and knowledge-based analysis techniques was developed for this experiment. The system essentially added acoustic confidence measures to the normal recognized input strings input to the existing MINDS-II dialog system. Specifically, the MINDS-II system was input with the words flagged inaccurate 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 should enhance performance if complementary sets of misrecognitions are discriminated using the two knowledge sources together.

Before we could evaluate whether the conjoined system could better detect misrecognized input than either component system alone, it was necessary to determine how to combine the knowledge sources. To develop decision rules, we analyzed the relative strengths and weaknesses, or the types of errors associated with each approach. We know that the MINDS system can accurately detect semantically or contextually inappropriate misrecognitions, but not those word substitutions that are consistent with prevailing constraints. The higher-level analyses are adept at detecting and separating many confusible content words, particularly those that can be distinguished on semantic grounds or made contextually distinct. Also MINDS has heuristics to recognize mid-utterance corrections and restarts and does not in-

accurately flag correctly recognized words.

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. Experiments 1 and 2 indicate that many of the words that cannot be reliably discriminated are high frequency and often short words. The acoustic error patterns are illustrated in Tables 3 and 1. The same *D'* analysis on the same test set is presented for the semantic/pragmatic/discourse system in Table 4.

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

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

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

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.

5.1. Decision Rules for Combining Knowledge Sources

Given these sets of results and the error patterns associated with each technique, we decided to begin experimenting with the following decision rule for combining the two knowledge sources. The rule has two parts. First, if the semantic/discourse module rejects a word string (or phrase) and decides it is misrecognized, the system will reject the string regardless of its probability correct. Second, if the acoustic module rejects a word string the system will reject it regardless of the semantic evaluation. In this way, the false rejection rate (rejecting correctly recognized words) should nether increase nor decrease, but the system does stand to significantly decrease the inaccurate acceptance rate.

5.2. Results of Conjoining 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 using the same ARPA ATIS test set previously employed. The test set contained approximately 5,000 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. A modified version of the SPHINX-I recognizer was used 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 was to detect the misrecognized word substitutions.

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

Table 5: Error Rate Reductions from Misrecognition Detection

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 test set using a modified SPHINX-I discrete HMM-based recognizer, 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 misrecognition detection performance relative to each of the individual systems. The conbined system decreased the number of missed (incorrectly accepted) misrecognized words by 41.3% and 34.1%, respectively, relative to the baseline performance of the acoustic and semantic/discourse techniques. The percentage of misclassified, correctly recognized words in the conjoined system did not increase.

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 meaningful 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 equipped with algorithms to detect mid-utterance corrections and restarts, those phenomena were readily flagged. Similarly, the semantic/pragmatic module detected misrecognitions caused by word confusibility in content words and identified inconsistent, misrecognized input. Both modules tended to confuse contractions with their associated expansions and had difficulty 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 6: D' Results for Conjoined Knowledge Sources

Given these results, and the success of this initial investigation, we now plan to more thoroughly integrate the acoustic confidence module with the MINDS-II system and fold them into the decoding search. 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 system more information to reason upon, may enhance its ability to detect misrecognized words and enables the system to analyze and evaluate those words with low probability acoustic confidence scores to perhaps decrease the number of false rejections. We also plan to investigate more sophisticated decision rules for conjoining knowledge sources. The initial decision rule employed in this study needs to be improved. Finally, we are investigating methods that use normalized acoustic scores and confidence measures during decoding and during the highly constrained, low perplexity re-recognition decoding search.

REFERENCES

1. 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.
2. 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.
3. 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.
4. Asadi, A., Schwartz, R., Makhoul, J., "Automatic Modeling for Adding New Words to a Large-Vocabulary Continuous Speech Recognition System", *ICASSP-91*, 1991, pp. 305-308.
5. 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.
6. Ward, W.H., "Modelling Non-Verbal Sounds for Speech Recognition", *Proceedings of the DARPA Speech and Natural Language Workshop*, October 1989.
7. 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.
8. 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.
9. 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.
10. Ward, W., "Understanding Spontaneous Speech: The PHOENIX System", *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1991, pp. .
11. 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.

*Different HMM based recognition systems have different overall error rates. Generally, the greater the error rate, the smaller the percentage of errors that are semantically consistent.