Recognition Confidence Scoring for Use in Speech Understanding Systems*

Timothy J. Hazen, Theresa Burianek, Joseph Polifroni and Stephanie Seneff

Spoken Language Systems Group
MIT Laboratory for Computer Science
545 Technology Square
Cambridge, MA 02139 USA
{hazen,tkburian,joe,seneff}@sls.lcs.mit.edu

ABSTRACT

In this paper we present an approach to recognition confidence scoring and a method for integrating confidence scores into the understanding and dialogue components of a speech understanding system. The system uses a multi-tiered approach where confidence scores are computed at the phonetic, word, and utterance levels. The scores are produced by extracting confidence features from the computation of the recognition hypotheses and processing these features using an accept/reject classifier for word and utterance hypotheses. The output of the confidence classifiers can then be incorporated into the parsing mechanism of the language understanding component. To evaluate the system, experiments were conducted using the JUPITER weather information system. Evaluation was performed at the understanding level using key-value pair concept error rate as the evaluation metric. When confidence scores were integrated into the understanding component of the system, the concept error rate was reduced by over 35%.

1. INTRODUCTION

The Spoken Language Systems Group conducts research leading to the development of conversational systems for human-machine interaction. These systems must not only recognize the words which are spoken by a user but also understand the user's query and respond accordingly. To achieve this goal, accurate automatic speech recognition is a necessity. The presence of incorrectly recognized words may cause the system to misunderstand a user's request, possibly resulting in the execution of an undesirable action.

Unfortunately today's speech recognition technology is far from perfect and errors in recognition must be expected. For example, let us consider the performance of the JUPITER weather information system [14]. On a randomly selected test set of 2388 utterances, the recognizer for JUPITER achieves a word error rate of 19.1%. On utterances which contain no out-of-vocabulary words and are clean of other artifacts that make recognition difficult (i.e., background noise, partial words, etc.) the error rate is only 9.9%. However, these "clean" utterances constitute only 75% of the test data. The error rate on the remaining 25% of the data is over 50%! It is this type of performance which motivates the development of confidence scoring techniques.

Because recognition errors can not yet be avoided, it alternatively becomes desirable for a system to be able to detect when recognition errors have occurred and take appropriate actions to recover from these errors. To provide an example, suppose a user asks JUPITER the following question:

what is the forecast for paramus park new jersey

As it happens, the JUPITER speech recognizer does not have the word *paramus* in its vocabulary. As such, the recognizer will provide its best guess using the words it knows. Thus, it might hypothesize the following query:

what is the forecast for **paris** *park new jersey*

Using confidence scoring techniques JUPITER should be able to determine that the word **paris** was not a reliable hypothesis. It could then mark this word as a potentially misrecognized word when passing the utterance on to the understanding component of the system. At that point the understanding component would need to be able to determine that the user is looking for the forecast for some place in New Jersey, but that the name of the place was misrecognized. Using this information the system could then prompt the user with the list of places in New Jersey for which it knows forecasts. The system might also prompt the user to spell the name of the city and learn it for future use.

To develop a system capable of the actions described above, two specific research goals must be addressed. First, a recognition confidence scoring technique which accurately determines when a recognizer's output hypothesis is reliable or unreliable must be developed. Second, confidence scores must be integrated into the back-end components of the system (e.g., language understanding and dialogue modeling) thereby enabling these components to make an informed decision about the action that should be taken when a confidence score indicates that a hypothesis may be incorrect. It is these two goals that our research strives to address. In this paper, we will present the details of our approach to this problem and present experimental results demonstrating the capabilities of our techniques.

*This research was supported by DARPA under contract N66001-99-1-8904, monitored through Naval Command, Control and Ocean Surveillance Center.

2. RECOGNITION CONFIDENCE SCORING

Overview

An accurate method for determining confidence scores for the speech recognition process must take into account two primary difficulties inherent in typical speech recognition systems. First, the models used in the recognition process may be inadequate, for any number of reasons, for discrimination between competing hypotheses. Second, recognizers are typically developed for *closed set* recognition (e.g., recognition using a pre-determined fixed vocabulary) and are thus not entirely appropriate for *open set* recognition problems where unknown words, partial words, and non-speech noises may corrupt the input.

Thus, an accurate confidence scoring technique should take into account the various factors which can contribute to misrecognitions. First, the scoring technique must be able to determine whether or not the recognizer has many competing hypotheses which could cause confusions. Recognition errors are less likely to occur when one hypothesis easily outscores all other competing hypotheses. Likewise, errors are far more likely to occur when multiple competing hypotheses all have similar scores [6]. Second, the recognizer must be able to determine if the input speech is actually a good fit to the underlying models used by the system, regardless of the relative scores of the competing hypotheses. Errors are more likely when there is a poor fit between the input test data and the training data. This can be the case when unknown words or non-speech sounds are present in the input data.

To attack this problem we utilize a technique where confidence scores are computed based on a set of confidence measures extracted from the computations performed during the recognition process [2, 9, 12]. For each recognition hypothesis, a set of confidence measures are computed and combined together into a confidence feature vector. The features which are utilized are chosen because, either by themselves or in conjunction with other features, they can be shown to be correlated with the correctness of a recognition hypothesis. The feature vectors for each particular hypothesis are then passed through a confidence scoring model which produces a single confidence score based on the entire feature vector. This score can then be evaluated by an accept/reject classifier which produces an accept/reject decision for the hypothesis. This approach is utilized in our work for both utterance level and word level confidence scores.

Phonetic Level Scoring

Many confidence scoring techniques focus on an examination of the scores produced by the recognizer's acoustic models at the phonetic level. Because the raw acoustic scores are usually not particularly useful as confidence measures when used by themselves [1], methods for normalizing these scores are typically employed [3, 8, 13]. In this work all of the acoustic scores produced at the phonetic level are normalized against a *catch-all* model. The normalization of the acoustic score does not affect the outcome of the recognition search but does allow the score produced for each phone to act as a phonetic level confidence feature. Mathematically, the phonetic level confidence score for a hypothesized phone u given an acoustic observation, $\vec{x}$, is:

$$c(u|\vec{x}) = \log \frac{p(\vec{x}|u)}{p(\vec{x})} \quad (1)$$

This normalization process produces a score which is zero-centered with respect to the log of $p(\vec{x})$, allowing the scores to be consistent across different observations. In practice, the *catch-all* model that is used is an approximation of the $p(\vec{x})$ model that would result from the weighted summation of the $p(\vec{x}|u)$ models over all u [7]. In this work, the individual phonetic scores are never used as independent confidence scores. However, they are used to help generate word and utterance level features. All references to *acoustic scores* in the remainder of this paper refer to the normalized acoustic scores described above.

Utterance Level Features

For each utterance a single confidence feature is constructed from a set of utterance level features extracted from the recognizer. For this work 15 different features which have been observed to provide information about the correctness of an utterance hypothesis were utilized. These features, as computed for each utterance, are:

1. **Top-Choice Total Score:** The total score from all models (i.e., the acoustic, language, and pronunciation models) for the top-choice hypothesis.

2. **Top-Choice Average Score:** The average score per word from all models for the top-choice hypothesis.

3. **Top-Choice Total N-gram Score:** The total score of the N-gram model for the top-choice hypothesis.

4. **Top-Choice Average N-gram Score:** The average score per word of the N-gram model for the top-choice hypothesis.

5. **Top-Choice Total Acoustic Score:** The total acoustic score summed over all acoustic observations for the top-choice hypothesis.

6. **Top-Choice Average Acoustic Score:** The average acoustic score per acoustic observation for the top-choice hypothesis.

7. **Total Score Drop:** The drop in the total score between the top hypothesis and the second hypothesis in the N-best list.

8. **Acoustic Score Drop:** The drop in the total acoustic score between the top hypothesis and the second hypothesis in the N-best list.

9. **Lexical Score Drop:** The drop in the total N-gram score between the top hypothesis and the second hypothesis in the N-best list.

10. **Top-Choice Average N-best Purity:** The average N-best purity of all words in the top-choice hypothesis. The N-best purity for a hypothesized word is the fraction of N-best hypotheses in which that particular hypothesized word appears in the same location in the sentence.

11. **Top-Choice High N-best Purity:** The fraction of words in the top-choice hypothesis which have an N-best purity of greater than one half.

12. **Average N-best Purity:** The average N-best purity of all words in all of the N-best list hypothesis.

13. **High N-best Purity:** The percentage of words across all N-best list hypotheses which have an N-best purity of greater than one half.

14. **Number of N-best Hypotheses:** The number of sentence hypotheses in the N-best list. This number is usually its maximum value of ten but can be less if fewer than ten hypotheses are left after the search prunes away highly unlikely hypotheses.

15. **Top-Choice Number of Words:** The number of hypothesized words in the top-choice hypothesis.

Word Level Features

For each hypothesized word, a set of word level features are extracted from the recognizer to create a confidence feature vector. For this work 10 different features, which have been observed to provide information about the correctness of a word hypothesis, were utilized. These features are:

1. **Mean Acoustic Score:** The mean log likelihood acoustic score across all acoustic observations in the word hypothesis.

2. **Mean Acoustic Likelihood Score:** The mean of the acoustic likelihood scores (not the *log* scores) across all acoustic observations in the word hypothesis.

3. **Minimum Acoustic Score:** The minimum log likelihood score across all acoustic observations in the word hypothesis.

4. **Acoustic Score Standard Deviation:** The standard deviation of the log likelihood acoustic scores across all acoustic observations in the word hypothesis.

5. **Mean Difference From Maximum Score:** The average difference, across all acoustic observations in the word hypothesis, between the acoustic score of a hypothesized phonetic unit and the acoustic score of highest scoring phonetic unit for the same observation.

6. **Mean Catch-All Score:** Mean score of the catch-all model across all observations in the word hypothesis.

7. **Number of Acoustic Observations:** The number of acoustic observations within the word hypothesis.

8. **N-best Purity:** The fraction of the N-best hypotheses in which the hypothesized word appears in the same position in the utterance.

9. **Number of N-best:** The number of sentence level N-best hypotheses generated by the recognizer.

10. **Utterance Score:** The utterance confidence score generated from the utterance features described above.

Classifier Training

The Training Data: To train the confidence scoring mechanism and the accept/reject classifier, a set of training data must be used which is independent of the training data used to train the recognizer. The independence is required to insure that the confidence scoring mechanism accurately predicts the recognizer's performances on *unseen* data. In our experiments, which were conducted using the JUPITER system, the confidence training data consists of 2506 JUPITER utterances. Each utterance is passed through the recognizer and then the N-best hypotheses (where N = 10) which are produced by the recognizer are used to train the confidence scoring mechanism. For word confidence scoring only the words in the top-choice hypothesis are used for training.

Data Labeling: The first step in the training process is to label the data. Each training token must be labeled either as *correct* or *incorrect*. The *correct* label is for tokens which should be accepted by the classifier, while the *incorrect* label is for tokens which should be rejected. This step must be taken for both the word and utterance level classifiers. In both cases, each *correct/incorrect* label is associated with the confidence feature vector extracted from the recognizer for that hypothesis.

For word level scoring the labeling scheme is obvious. Correctly hypothesized words are labeled as *correct* and incorrectly hypothesized words are labeled as *incorrect*.

For utterance level scoring the concept of correctness is not as clear. We have elected to use a set of heuristics to define the labels of *correct* and *incorrect* such that only utterances which the recognizer has extreme difficulties recognizing will be marked as incorrect. In this labeling scheme, we mark utterances in which the correct orthography is one of the top four sentence hypotheses as *correct*. Utterances in which at least two out of every three words in the top-choice hypothesis are correctly recognized are also marked as *correct*. All other utterances are labeled as *incorrect*.

The Classifier Model: The same confidence scoring technique is used for both word and utterance level confidence scoring. To produce a single confidence score for a hypothesis, a simple linear discrimination projection vector is trained. This projection vector reduces the multi-dimensional confidence feature vector from the hypothesis down to a single confidence score. Mathematically this is expressed as

$$r = \vec{p}^{\,\mathrm{T}} \vec{f} \tag{2}$$

where $\vec{f}$ is the feature vector, $\vec{p}$ is the projection vector, and r is the raw confidence score.

Because the raw confidence score r is simply a linear combination of a set of features, the score has no probabilistic meaning. Ideally, we prefer to generate scores which have a probabilistic meaning in order to make these scores more compatible with other probabilistic components of our entire system. To this end, a probabilistic confidence score based on maximum *a posteriori* probability (MAP) classification is created using the following expression:

$$c = \log\left(\frac{\mathrm{p}(r|\mathrm{correct})\mathrm{P}(\mathrm{correct})}{\mathrm{p}(r|\mathrm{incorrect})\mathrm{P}(\mathrm{incorrect})}\right) - t \qquad (3)$$

In this expression, $\mathrm{p}(r|\mathrm{correct})$ and $\mathrm{p}(r|\mathrm{incorrect})$ are Gaussian density functions for r for correct and incorrect tokens, $\mathrm{P}(\mathrm{correct})$ and $\mathrm{P}(\mathrm{incorrect})$ are *a priori* probabilities of observing correct or incorrect tokens, and c is the final probabilistic confidence score expressed in the log domain. Note that a constant decision threshold t is applied to the score to set the accept/reject decision threshold to zero. Thus, after the decision threshold t is subtracted, a negative score for c results in a *rejection* while a non-negative score results in an *acceptance*.

The Training Method: The projection vector $\vec{p}$ is trained using a *minimum classification error* (MCE) training technique. In this technique the projection vector $\vec{p}$ is first initialized using Fisher Linear Discriminant analysis. After the initialization of $\vec{p}$, a simple hill-climbing algorithm iterates through each dimension in $\vec{p}$ adjusting its values to minimize the classification error rate on the training data. The optimization continues until a local minimum in error rate is achieved. The Gaussian density parameters of the classifier model are trained from the raw scores generated after applying $\vec{p}$ to the feature vectors in the training set.

The threshold t is determined by setting the operating point of the system to a desired location on the *receiver-operator characteristic* (ROC) curve. For the utterance level scores, the threshold is set such that 98% of the utterances which are labeled as correct are accepted. This threshold is chosen to insure a high detection rate which discourages false rejections. For words, the minimum classification error rate is chosen as the desired operating point.

Experiment Test Conditions

To test the confidence scoring techniques, a test set of 2388 JUPITER utterances is utilized. For recognition we utilize the SUMMIT speech recognition system [4] as trained specifically for the JUPITER domain [5]. The recognizer is trained from over 70,000 utterances collected from live telephone calls to our publicly available system. The recognizer's vocabulary has 2005 words. As discussed in the introduction, the recognizer achieved a word error rate of 19.1% on this test set.

Utterance Level Experimental Results

The goal of utterance level confidence scoring is to reject utterances with which the recognizer has extreme difficulty. With this in mind the utterance scoring mechanism rejected 13% of the utterances in the test set. The word error rate on this 13% of the data was over 100% (e.g., there were more errors than actual words in the reference orthographies). Closer examination reveals that only 27% of the reference words in the orthography were actually recognized correctly and that both substitution errors and insertion errors happened more frequently than correct recognitions. By comparison, the word error rate on the 87% of the utterances that were accepted was 14%. These results indicate that the utterance level confidence scoring mechanism performs its job as intended.

Word Level Experimental Results

To evaluate word level confidence scoring, we have chosen to use the error rate of the accept/reject classifier. Using this evaluation metric, an error occurs if the classifier accepts a misrecognized word or rejects a correctly recognized word. This error rate is directly related to a recognition metric we refer to as the *hypothesized word error rate* (HWER). The hypothesized word error rate is expressed as follows:

$$\mathrm{HWER} = \frac{(\#\text{ of substitutions}) + (\#\text{ of insertions})}{\#\text{ of reference words}} \qquad (4)$$

The HWER differs from the standard word error rate (WER) in that it neglects deletion errors. This metric is related to the accept/reject error rate because the accept/reject classifier can only operate on words which are actually present in the hypothesis. At present the confidence scoring technique has no ability to express the confidence that a word may have been deleted. The relationship between the accept/reject error rate and the HWER results from the fact the HWER acts as an upper bound on the accept/reject error rate. This can be achieved by instructing the classifier to accept all word hypotheses. This assumes that the HWER is less than 50%. In cases where the HWER is actually greater than 50% the upper bound is based on a system which instead rejects all hypothesized words. With this in mind, the goal is to achieve an accept/reject error rate which improves upon this upper bound. The system which simply accepts (or rejects) all words will be referred to as the *baseline* system against which the accept/reject classifier is compared.

Table 1 examines the accept/reject classification error rate under three conditions: (1) the baseline system, (2) a classifier using each of the 10 word features on an individual basis, and (3) the system using the complete set of features with the MCE trained linear discriminant classifier. These results were computed over all hypothesized words from only utterances accepted by the utterance level classifier. As can be seen in the table, the individual features based solely on the acoustic scores do not perform particularly well by themselves. In fact, the mean log-likelihood acoustic score, which is the best of the acoustically-based confidence features, has an accept/reject error rate which is only 3% less than the baseline system (11.9% vs. 12.1%). By comparison, the utterance level score, which is the same for all words in any sentence hypothesis, yields a 7% improvement from the baseline (11.2% vs. 12.1%), and the N-best purity measure yields an 11% improvement (10.8% vs. 12.1%). By combining all of the features together an error rate reduc-

Test Condition or Feature	Accept/Reject Error Rate
Baseline (HWER)	12.1 %
# of N-best	12.1 %
Acoustic Score Std. Dev.	12.1 %
# of Acoustic Observations	12.1 %
Mean Catch-All Score	12.1 %
Minimum Acoustic Score	12.1 %
Mean Diff. from Max Score	12.0 %
Mean Acoustic Likelihood	11.9 %
Mean Acoustic Score (log)	11.7 %
Utterance Score	11.2 %
N-best Purity	10.8 %
Combined	9.4 %

Table 1: Accept/reject classification performance of word confidence scoring mechanism on accepted utterances when each feature is tested independently and when features are combined using linear combination with Minimum Classification Error training.

tion of 22% from the baseline can be achieved (9.4% vs. 12.1%).

Table 2 shows the performance of the classifier under two different constraints. First, the table shows the performance of the classifier when tested on accepted versus rejected utterances. When tested on accepted utterances the classifier is intended to detect as many misrecognized words as possible while maintaining a low false rejection rate. At this time our system does not actually examine utterances which have been rejected. However, rejected utterances could conceivably be scanned for important content words that are accepted based on their word confidence score. In the table the baseline system error rate for rejected utterances is the error rate when all hypothesized words are rejected. This results from the fact that 72.8% of the hypothesized words in the rejected utterance are incorrect. As can be seen in the table, the classifier shows a larger reduction in classifier error rate from the baseline on rejected utterances than it does on accepted utterances. This result indicates that the word confidence scoring technique can be useful for both accepted and rejected utterances, even though our system currently only applies it to accepted utterances.

Table 2 also shows the performance of the classifier when applied to all hypothesized words as compared to its application to only hypothesized words which are proper names of geographic locations. This analysis is useful because content words such as location names are typically more important to the correct understanding of an utterance than function words. The results indicate that the confidence scoring technique is more accurate when examining the performance on hypothesized location names than it is over all words in general. This result is very satisfying since it indicates that the confidence scoring technique works best on the words which are most important for understanding.

The performance of the accept/reject classifier can also be examined in several other interesting ways. When examining accepted utterances only, the system correctly rejects 51% of the incorrectly hypothesized words while only falsely rejecting 4% of correct words. These numbers improve to 54% and 3.5% when considering only words which are location names. Furthermore, across all utterances the combination of utterance and word level scoring correctly detects 72% of the errors introduced by unknown words and 85% of the errors introduced by non-lexical artifacts.

Utterances	Words	Accept/Reject Error Rate	
		Baseline	Classifier
All	All words	16.4 %	10.1 %
Accepted	All words	12.1 %	9.4 %
Rejected	All words	27.2 %	19.1 %
All	Locations	17.8 %	9.1 %
Accepted	Locations	12.9 %	8.7 %
Rejected	Locations	24.3 %	14.5 %

Table 2: Comparison of accept/reject classification performance of word confidence scoring mechanism over all utterances, accepted utterances only, and rejected utterances only when considering all hypothesized words versus geographic location words only.

3. INTEGRATING CONFIDENCE SCORES INTO UNDERSTANDING

Overview

While it is interesting to examine the results of the confidence scoring techniques in the context of recognition, the ultimate goal of this work is to improve the understanding accuracy of our conversational systems. To achieve this, we must integrate the recognition confidence scores into the language understanding component of the system. For language understanding we utilize the TINA natural language understanding system [11]. TINA utilizes a semantically-tagged context free grammar to parse each utterance. In cases where TINA is unable to generate a full parse, the system may back off to a robust (or partial) parse of the utterance. For utterances in which either a full or robust parse is found, a set of semantic concepts, represented as key-value pairs, can be extracted from the semantic information present in the parse tree. In our experiments, language understanding is evaluated by examining the *concept error rate* from the set of key-value pairs [10].

To integrate confidence scores into the understanding component a two-step process is utilized. First, if an utterance is rejected at the utterance level, the understanding component does not attempt to understand the utterance and assumes that no useful information for understanding can be extracted from the recognizer's output. In this case the system does not generate any key-value pairs. If the utterance is accepted, the second step is to create an N-best list which is augmented with confidence scores, and allow the natural language parser to try to interpret the utterance, given that some words may be misrecognized.

N-best List Augmentation

To handle word confidence scores, only a few modifications to the basic N-best list are required. First, the N-best list passed to the parser is augmented with confidence

N-best list without rejection:

what_is	6.13	the	5.48	forecast	6.88	for	5.43	**paris**	**-0.03**	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	4.47	**hyannis**	**-0.16**	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	5.12	**venice**	**-1.49**	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	4.28	**france**	**-1.76**	park	4.41	new_jersey	4.35

N-best list with hard rejection:

what_is	6.13	the	5.48	forecast	6.88	for	5.43	***reject***	**0.00**	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	4.47	***reject***	**0.00**	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	5.12	***reject***	**0.00**	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	4.28	***reject***	**0.00**	park	4.41	new_jersey	4.35

N-best list with optional rejection:

what_is	6.13	the	5.48	forecast	6.88	for	5.43	***reject***	**0.00**	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	5.43	**paris**	**-0.03**	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	4.47	***reject***	**0.00**	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	4.47	**hyannis**	**-0.16**	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	5.12	***reject***	**0.00**	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	5.12	**venice**	**-1.49**	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	4.28	***reject***	**0.00**	park	4.41	new_jersey	4.35
what_is	6.13	the	5.48	forecast	6.88	for	4.28	**france**	**-1.76**	park	4.41	new_jersey	4.35

Table 3: Example N-best lists augmented with confidence scores. The first list is the standard output from the recognizer. The second list shows how the rejected word alternative are added to the first list before being passed on to the understanding component.

scores. The first list in Table 3 shows an example N-best list augmented with confidence scores. Two different word rejection strategies can be applied to the initial N-best list. The second list in Table 3 shows the application of *hard rejection* to the N-best list. In this case, any word with a confidence of less than zero is replaced with a rejected word marker which receives the neutral score of zero. The third list in Table 3 shows the application of *optional rejection*. This list is essentially the combination of the first two lists. Using optional rejection, poorly scoring words are retained in the final N-best list but must compete with the rejected word markers they generate, which have a higher score.

Word Graph Search

Within TINA, the incoming N-best list is collapsed into a word graph. Each arc in the word graph is augmented with a score for its respective word. Before the implementation of word level confidence scores, a heuristic word scoring method was utilized which generated scores based on the number of N-best hypotheses each word appeared in and the rank of those N-best hypotheses [7]. In the new version of the system, each arc in the word graph is augmented with the word-level confidence scores generated from the recognizer.

The parser performs a beam search through the graph combining the word scores with trained linguistic probabilities to generate a total score for each parse theory. From a ranked list of parse theories extracted from the word graph search, TINA selects the highest scoring theory that produces a full parse. If no path through the word graph can be found that generates a full parse then the system selects the highest scoring robust parse. The disadvantage of this approach is that it has the possibility of selecting any word sequence through the word graph in order to find a sentence that parses, even if one or more words in the hypothesis are highly likely to be misrecognitions. When the input N-best list is augmented with word rejections, the resulting word graph allows the parser the option of selecting rejected words instead of poorly scoring words.

Grammar Augmentation

In order to utilize an N-best list containing rejected words, the grammar must be augmented to accept rejected words in specific contexts. For our experiments with JUPITER, only two modifications to the grammar were made. First, the grammar was adjusted to allow rejected words to be parsed as *unknown city names* in sentence contexts where the rejected word was almost certainly a city name. In the example in Table 3, for example, the word sequence "*reject* park" would be parsed as an unknown city name. This adjustment complemented an existing parsing mechanism which allowed unknown words (i.e., words not in the vocabulary of the grammar) to parse in a similar fashion.

The second adjustment to the grammar was to allow rejected words appearing anywhere in the sentence to be skipped when the parser is attempting to find a robust parse. This allows the parser to concentrate on only the portions of the utterance which were recognized with high confidence. This modification is especially useful for eliminating problems that result from spurious sounds or speech at the beginning and/or end of an utterance.

Experimental Results

To examine the effects of confidence scoring on language understanding, the JUPITER system can be evaluated on the test data under five different conditions: (1) using the original system which did not utilize word confidence scores, (2) using the new system which utilizes word confidence scores

Experimental Conditions	Error Rates (%)			
	Sub.	Ins.	Del.	Total
Original system	1.9	20.2	6.4	28.5
New system w/o reject.	2.1	18.2	6.1	26.3
+ utterance rejection	1.8	12.7	7.1	21.7
+ optional word reject.	1.3	9.0	8.4	18.7
+ hard word rejection	1.0	7.2	10.5	18.6

Table 4: Understanding error rates as confidence scores and different levels of confidence rejection are added to the system.

but does not perform any rejection, (2) using the new system with utterance rejection, (3) using the new system with utterance rejection and optional word rejection, and (4) using the new system with utterance rejection and hard word rejection. As discussed earlier, these conditions are investigated using key-value pair concept error rate [10]. The results are shown in Table 4 in terms of substitution, insertion, deletion, and total error rates. For these experiments, a substitution error occurs when a test utterance has a key-value pair where the key matches a key-value pair in the correct answer, but the value in the pair is different. An insertion occurs when a key-value concept is erroneously inserted. Likewise, a deletion occurs when a key-value concept is erroneously deleted.

An examination of Table 4 yields several important observations. First, the new system using the probabilistic word confidence scores has an error rate which is 8% smaller than the error rate of the original system using the heuristic word scores. However, both the original and new systems suffer from excessive insertion errors when no rejection is utilized. This is primarily the result of the understanding component's aggressive effort to find a reasonable interpretation of an utterance from any of the hypotheses in the N-best list. Without rejection, the understanding component can latch onto any hypothesis which produces a parse regardless of whether or not the recognizer is confident in the hypothesis. This generally produces the correct answer when the user is cooperative, speaks clearly and stays within domain. However, this approach yields many insertions when the utterance is out of domain, has unknown words, or has artifacts which cause difficulty for the recognizer.

Next, when utterance level rejection is added, the insertion error rate is reduced from 18.2% to 12.7% while the deletion error rate is only increased from 6.1% to 7.1%. In other words, the use of utterance rejection removes 5.5 insertion errors for every deletion error that is added. This translates into a relative error rate reduction of 17%.

Next, the addition of word rejection to utterance rejection produces another significant improvement in the total error rate. While the total error rates for optional word rejection versus hard word rejection are virtually the same, the nature of the underlying errors is slightly different. Using optional word rejection, the insertion error rate remains higher than the deletion error rate. However, hard word rejection produces a result where deletions outnumber insertions. The relative desirability of each method would thus be dependent on whether or not insertion errors are more harmful to the user's interaction with the system than deletions. The addition of word rejection reduces the error rate by 14% from the system using utterance rejection only. Overall, the use of utterance and word confidence scores and rejection within the understanding component of the system reduces the understanding error rate by 35% from 28.5% to 18.6%.

4. DIALOGUE MODELING ISSUES

At this time, we are only just beginning to consider the dialogue modeling issues involved in utilizing the confidence scoring techniques that we have presented here. At the present time, only two dialogue actions have been implemented which take advantage of the confidence scoring capability. The first action is the response the dialogue manager generates when an entire utterance has been rejected. Under this circumstance, the system only knows that the recognizer could not produce any viable hypothesis for the input utterance. Since very little else is known, the system cannot provide an informed response to the user about its failure to understand the utterance. When this happens the system simply informs the user that a misunderstanding has occurred and then provides a generic help message which will hopefully guide the user in the right direction. If the user continues to produce utterances that are rejected, the system will generate a new help message for each new rejected utterance.

The second action that the system is currently capable of performing is invoked when the system is able to recognize that the user is asking for a city but the city name has been rejected. In this event, the system first informs the user that it isn't sure what city the user is asking about. After that, it can generate one of two different follow-on responses to help guide the user. First, if the user provided a state or country name in addition to the city name, the system will provide a list of cities that it knows about in that state or country and request that the user select one. This helps guide the user towards city names that are in the vocabulary of the recognizer. Second, if the user did not mention a state or country name, the system asks the user to specify a state of country name in order to help guide the user hierarchically towards a city that the system knows about. An example dialogue which illustrates some of these actions is shown in Table 5.

5. CONCLUSION

In this paper we have presented a multi-tiered confidence scoring approach which is able to produce confidence scores at the phonetic, word and utterance levels. These techniques have proven to be especially useful when the speech signal contains unknown words or non-lexical artifacts. This is evidenced by the fact that the system correctly detects 72% of the errors introduced by unknown words and 85% of the errors introduced by non-lexical artifacts. By integrating the word and utterance confidence scoring techniques into the understanding component of the JUPITER weather information system, we were able to reduce the understanding error rate of the system by 35%.

Turn 1:	User:	What is the forecast for Paramus Park, New Jersey?
	Recognizer:	*what is the forecast for* ***reject*** *park new jersey*
	System:	I may not know the city you're asking about. I know of the following places in New Jersey: Atlantic City, Camden, Newark, Piscataway, Princeton, and Trenton. Please select one.
Turn 2:	User:	Piscataway.
	Recognizer:	*piscataway*
	System:	In Piscataway today, cloudy, 70 percent chance of rain, high 60 low 42. Can I help you with something else?
Turn 3:	User:	Will the Yankees game against the Blue Jays be rained out tonight?
	Recognizer:	**utterance rejected**
	System:	Sorry I'm not sure what you said.
Turn 4:	User:	I asked if the Yankees game will be rained out tonight.
	Recognizer:	**utterance rejected**
	System:	I'm sorry. I may not know the city you're asking about. Try asking for the state or country. For example: "What cities do you know about in Massachusetts?"
Turn 5:	User:	I guess you can't answer that. Good bye.
	Recognizer:	*i guess you* ***reject* *reject* *reject*** *good bye*
	System:	Thank you for calling. Have a nice day!

Table 5: Example dialogue demonstrating actions taken when words and utterances are rejected.

6. FUTURE WORK

As part of our continuing research on the topic, we hope to improve our use of confidence scoring in several ways. First, we would like to expand the use of rejected words in our augmented grammars. Because contextual information can be a powerful predictor of the semantic class of misunderstood words, we hope to be able to utilize the same technique we employed with city names to help predict the semantic class of other types of rejected words (such as weather conditions or dates).

Next, we hope to expand the set of dialogue actions that can take advantage of the confidence scores. Two possible dialogue actions that will be investigated are confirmation (i.e., "Did you say Boston?") and clarification (i.e., "Did you say Boston or Austin?"). Another possible action might be for the system to request the spelling of a location that the user is asking about which is not in the recognizer's vocabulary. In order to handle these actions, the current set of two confidence regions (*accept* and *reject*) must be expanded to include a third region of *uncertain*. The use of these new dialogue actions could especially help improve the nature of a dialogue at a time when the confidence in a recognition hypothesis is neither extremely high nor extremely low.

Finally, we wish to explore the use of these techniques across a wide variety of systems. We hope to discover the aspects of the techniques which work well across all domains and the aspects which are somewhat domain-dependent. Ultimately, a confidence scoring technique which is as domain independent as possible will be most useful for the rapid deployment of systems in new domains.

ACKNOWLEDGMENTS

The authors wish to acknowledge the contributions of Jim Glass, Christine Pao, Philipp Schmid, and Simo Kamppari, whose prior work on confidence modeling laid the foundation for the experiments presented in this paper.

REFERENCES

[1] Z. Bergen and W. Ward, "A senone based confidence measure for speech recognition," In *Proc. of Eurospeech*, Rhodes, 1997.

[2] L. Chase, "Word and acoustic confidence annotation for large vocabulary speech recognition," In *Proc. of Eurospeech*, Rhodes, 1997.

[3] S. Cox and S. Dasmahapatra, "A high-level approach to confidence estimation in speech recognition," In *Proc. of Eurospeech*, Budapest, 1999.

[4] J. Glass, J. Chang, and M. McCandless, "A probabilistic framework for feature-based speech recognition," In *Proc. of ICSLP*, Philadelphia, 1996.

[5] J. Glass, T. Hazen and L. Hetherington, "Real-time telephone-based speech recognition in the JUPITER domain," In *Proc. of ICASSP*, Phoenix, 1999.

[6] L. Hetherington, *A Characterization of the Problem of New, Out-of-Vocabulary Words in Continuous-Speech Recognition and Understanding*. PhD thesis, MIT, 1994.

[7] S. Kamppari, *Word and Phone Level Acoustic Confidence Scoring for Speech Understanding Systems*. Master's thesis, MIT, 1999.

[8] S. Kamppari and T. Hazen, "Word and phone level acoustic confidence scoring," In *Proc. of ICASSP*, Istanbul, 2000.

[9] C. Pao, P. Schmid, and J. Glass, "Confidence scoring for speech understanding," In *Proc. of ICSLP*, Sydney, 1998.

[10] J. Polifroni, *et al*, "Evaluation Methodology for a Telephone-based Conversational System,"" In *Proc. Int. Conf. on Language Resources and Evaluation*, Granada, Spain, 1998.

[11] S. Seneff, "TINA: A natural language system for spoken language applications," *Computational Linguistics*, vol. 18, no. 1, March 1992.

[12] M. Sui, H. Gish, and F. Richardson, "Improved estimation, evaluation and applications of confidence measures for speech recognition," In *Proc. of Eurospeech*, Rhodes, 1997.

[13] G. Williams and S. Renals, "Confidence measures derived from an acceptor HMM," In *Proc. of ICSLP*, Sydney, 1998.

[14] V. Zue, *et al*, "JUPITER: A telephone-based conversational interface for weather information," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 1, January 2000.