A COMPARISON AND COMBINATION OF METHODS FOR OOV WORD DETECTION AND WORD CONFIDENCE SCORING[1]

Timothy J. Hazen and Issam Bazzi

Spoken Language Systems Group
MIT Laboratory for Computer Science
Cambridge, Massachusetts 02139 USA
hazen@sls.lcs.mit.edu , issam@mit.edu

ABSTRACT

This paper examines an approach for combining two different methods for detecting errors in the output of a speech recognizer. The first method attempts to alleviate recognition errors by using an explicit model for detecting the presence of out-of-vocabulary (OOV) words. The second method identifies potentially misrecognized words from a set of confidence features extracted from the recognition process using a confidence scoring model. Since these two methods are inherently different, an approach which combines the techniques can provide significant advantages over either of the individual methods. In experiments in the JUPITER weather domain, we compare and contrast the two approaches and demonstrate the advantage of the combined approach. In comparison to either of the two individual approaches, the combined approach achieves over 25% fewer false acceptances of incorrectly recognized keywords (from 55% to 40%) at a 98% acceptance rate of correctly recognized keywords.

1. INTRODUCTION

The Spoken Language Systems Group conducts research leading to the development of conversational speech understanding 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. The success of such systems is heavily dependent on the ability of the speech recognition component to accurately recognize the words spoken by the user. 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. Under these circumstances it becomes desirable to develop methods which can identify when a recognizer's hypothesis is correct and when it may be in error. In order to create methods to accomplish this, it is important to understand the two primary deficiencies in a typical recognizer. 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 predetermined 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.

In this paper we examine two methods, which were previously developed in our group, to help detect and alleviate the presence of errors in speech recognition hypotheses. In the first method, an explicit out-of-vocabulary (OOV) word model is added into the model set of the recognizer in order to identify potential unknown words during recognition [1]. In the second method, the recognizer's hypotheses are post-processed with a confidence scoring model in order to identify hypothesized words which may be misrecognized [5]. Both methods attempt to identify the regions of an utterance where the recognizer cannot find reliable word hypotheses without harming the regions where the recognizer is performing correctly. However, because the modeling approaches are inherently different, they have different advantages and disadvantages. Under these circumstances a combination of the two methods might prove beneficial. In this paper we seek to compare and contrast the performance of the two individual methods. This paper also presents a method for combining the techniques and provides experimental results demonstrating the performance gains that can be obtained by the combined approach. Results are presented using the recognizer for the JUPITER weather information system [13].

2. MODELING OOV WORDS

In devising a technique for explicitly modeling OOV words during recognition, we start with a word-based recognizer with a fixed predefined vocabulary of words. To model OOV words, we create a *generic word model* which must allow for arbitrary phone sequences during recognition. One simple generic word model is a phone recognizer covering the set of all phonetic units in the language. Since this unit inventory can cover the phonetic sequences contained in all possible OOV words, it can be used as the basis for our generic word model.

To allow for OOV words, the word-based recognizer's vocabulary is augmented with an OOV word whose underlying model is the generic word model. Figure 1 shows how the word search space can be augmented with the generic word model. We simply allow the search to transition into the generic word model W_{OOV} at the completion of any word. When exiting W_{OOV}, the search is allowed to either end the utterance or enter any other word model, including the OOV word.

Entrance into the generic word model is controlled by two pa-

[1] This research was supported by DARPA under contract N66001-99-8904, monitored through the Naval Command, Control and Ocean Surveillance Center, and by the NSF (National Science Foundation) under Grant No. IRI-9618731.

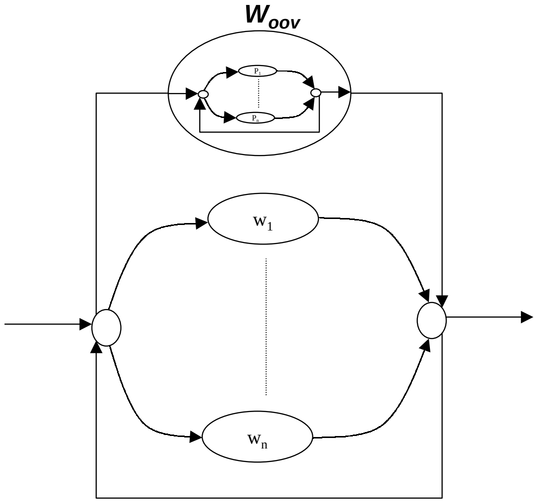

Fig. 1. The hybrid recognition configuration containing the OOV generic word model.

rameters during recognition. The first is an OOV cost, C_{OOV}. This cost is related to the probability of observing an OOV word and is used to balance the contribution of the OOV phone grammar to the overall score of the utterance. For our experiments we varied the value of C_{OOV} to quantify the behavior of the hybrid recognizer. The second parameter is simply the language model. The language model of the hybrid recognizer remains word-based, but now includes an entry for the OOV word. Since the OOV word is part of the vocabulary, the n-gram grammar treats the OOV word just like any other word in the vocabulary.

Augmenting the word recognizer with the generic word model shown in Figure 1 is somewhat similar to using filler (or garbage) models for word-spotting [10, 11]. However, there are two key distinctions which differentiate our approach from using filler models for word-spotting. First, the entire word vocabulary is used in the search, whereas the generic word is intended only to cover OOV words. The second distinction is that accurate sub-word recognition is important for our OOV model since it is possible to use its output for a second stage of processing to identify the pronunciation (and possibly the spelling) of the OOV word. In contrast, word spotters typically make no use of the output of the filler models.

3. WORD CONFIDENCE SCORING

In our system, word confidence scores are computed as a post-processing stage after recognition [7, 9, 12]. To obtain the confidence scores we begin by extracting a set of confidence measures for each word from the computations performed during the recognition process. In our system ten different confidence measures are computed. These include such measurements as the average normalized log-likelihood acoustic model score over all acoustic observations in a word, the minimum normalized log-likelihood acoustic model score for a word, the fraction of the N-best utterance hypotheses in which a hypothesized word appears, etc. These features are concatenated into a single confidence feature vector.

The feature vector for each individual word hypothesis is then evaluated using a confidence scoring model which produces a single confidence score based on the entire feature vector. To produce a confidence score for a word from the confidence feature vector, a simple linear discrimination projection vector is trained. This projection vector reduces the multi-dimensional confidence feature vector for the hypothesis down to a single confidence score. Mathematically this is expressed as

$$c = \vec{p}^{\,\mathrm{T}} \vec{f} \qquad (1)$$

where $\vec{f}$ is the feature vector, $\vec{p}$ is the projection vector, and c is the raw confidence score. A threshold on this score can be set to produce an accept/reject decision for the word hypothesis. In our experiments, this threshold is varied to adjust the balance between false acceptances of misrecognized words and false rejections of correctly recognized words. In [6], we describe how the raw score can be converted into a probabilistic score which can be used in later processing by the language understanding and dialogue components of the system.

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 MCE algorithm iterates through each dimension in $\vec{p}$ adjusting its values to minimize the accept/reject classification error rate on the training data. The optimization continues until a local minimum in error rate is achieved. Though this discriminatively trained projection vector approach is quite simple, it has performed quite well for us. Never the less, future work may attempt to use a more powerful accept/reject classifier such as a neural network [9, 12].

4. COMBINING OOV WORD DETECTION AND CONFIDENCE SCORING

In speech recognition research, it has been discovered that combining the outputs of different classifiers and/or recognizers can improve recognition accuracy and robustness [2, 4, 8]. These results are most compelling when the different classifiers utilize different observation measurements or modeling approaches but achieve similar results. Under these circumstances, the expected gain from combining the different classifiers is the greatest. This was the motivation for attempting to combine our two distinctly different methods for detecting recognition errors.

Our OOV word modeling approach operates during the recognition search process by allowing the recognizer itself to hypothesize a generic OOV word model as an alternative to a known word. On the other hand, our confidence scoring approach is applied as a post-processing technique after the recognition search is already complete. A natural way to combine both methods is to enable OOV word detection during recognition and then utilize confidence scoring on the hypothesized known words (excluding the OOV word hypotheses) after recognition is complete.

Using this two-stage combined approach there are two opportunities for the system to detect potential errors. During the recognition stage the OOV word detection approach replaces potential misrecognitions with unknown word markers. In the post processing stage, the confidence scoring module examines the remaining word hypotheses which are in-vocabulary and rejects those word hypotheses in which it has low confidence.

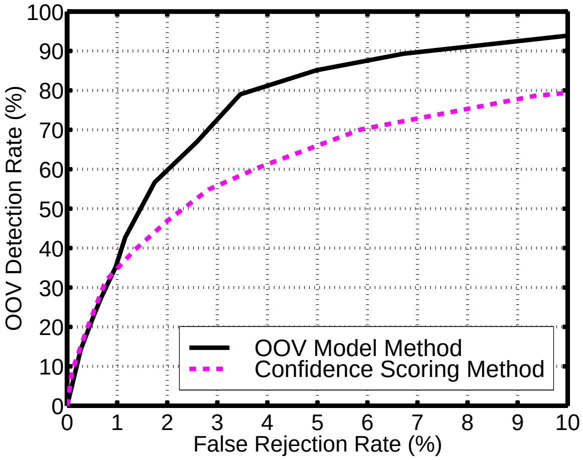

Fig. 2. Comparison of the rejection rate of errors caused by OOV words versus the false rejection rate of correctly recognized words.

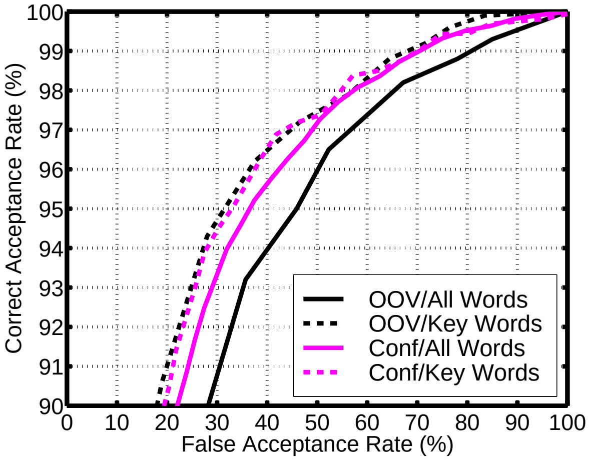

Fig. 3. ROC curves for OOV word detection and confidence scoring methods evaluated on all words and on keywords only.

5. EXPERIMENTS & RESULTS

5.1. Experimental Setup

Experiments presented here utilize the recognizer for the JUPITER weather information domain [3]. This recognizer utilizes a set of context-dependent diphone acoustic models, whose feature representation was based on the first 14 MFCCs averaged over 8 regions near hypothesized phonetic boundaries. Diphones were modeled using diagonal Gaussians with a maximum of 50 mixtures per model. The word lexicon consisted of a total of 2,009 words, many of which have multiple pronunciations. Word class trigram language models were used at the word-level. Phone-level bigrams were used for the internal phone transitions in the OOV model. The training set used for these experiments consists of 88,755 utterances used to train both the acoustic and the language models. The test set consists of 2,388 spontaneous utterances collected by JUPITER, 13% of which contain OOV words. On this test set the baseline recognizer has a word error rate of 21.6%.

In our experiments, we first examine the capability of the OOV detection method and the confidence scoring method on the task of detecting errors caused by unknown words. Second we compare the two methods on the task of detecting recognition errors in general. Finally, we examine the method for combining the two approaches on the task of keyword recognition error detection.

5.2. Detecting OOV Words

The purpose of the OOV word detection model is to detect the presence of OOV words without harming the recognition accuracy on correctly recognized known words. Similarly, it is hoped that the confidence scoring module will reject word hypotheses when the actual word is an unknown word without absorbing false rejections of correctly recognized known words. The performance of the two methods on the task of OOV word detection is shown in Figure 2. In this figure OOV word detection (i.e., the rejection of word hypothesis errors caused by unknown words) is plotted against the false rejection rate of correctly recognized words. As can be seen in the figure the OOV detection method performs better at the task of detecting errors caused by OOV words than the confidence scoring method. This is not surprising considering that the OOV detection method is designed specifically for this task while the confidence scoring method is designed for the more general task of detecting *any* type of recognition error (including substitution of known words and insertions).

5.3. Detecting Recognition Errors

As mentioned earlier, the confidence scoring model is designed to be a generic detector of recognition errors. Its focus is not specifically on the detection of errors caused by unknown words, as examined in the previous section. To test this capability, we can examine the receiver-operator characteristic (ROC) curve of the system. The ROC curve measures the relationship between the percentage of correctly recognized words which are accepted (i.e., the correct acceptance rate) against the percentage of incorrectly recognized word which are accepted (i.e., the false acceptance rate). Ideally we'd like to minimize the false acceptance rate without harming the correct acceptance rate.

Figure 3 shows four different ROC curves. The solid ROC curves show the OOV detection method and the confidence scoring method when applied to all words hypothesized by the recognizer. These lines indicate that the confidence scoring method has a better ROC curve than the OOV detection method when applied to all hypothesized words. This result is not surprising considering that the confidence scoring method was specifically designed for this task, while the OOV detection method was designed specifically for detecting errors caused by OOV words.

However, the dashed lines in Figure 3 show the ROC curves for the two methods when only examining certain keywords which are important for correct understanding. These keywords are from a list of 937 proper names of geographic locations known by the recognizer. For this test the two methods perform almost identically. The fact that the OOV detection method works much better on this keyword evaluation than it did on the evaluation using all words is also not surprising. Many of the out-of-vocabulary words that appear in the JUPITER task are proper names of locations. Because of language modeling constraints it is relatively common for the baseline recognizer to substitute a known location name for an out-of-vocabulary location name.

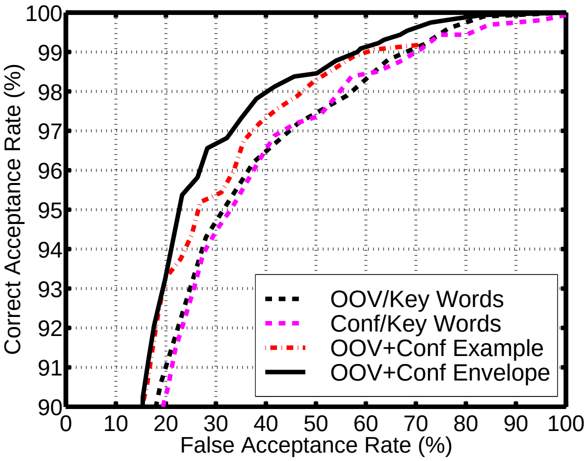

Fig. 4. ROC curves on hypothesized keywords only using the OOV word detection and confidence scoring methods as well as a combined approach.

5.4. The Combined Approach

Figure 4 shows the ROC curves for keyword detection for our original two methods plus the combined method. In the combined approach, the OOV modeling component is first fixed at a particular operating point before confidence scoring is utilized. The dashed and dotted line on the figure shows one example ROC curve for the combined approach for one initial OOV modeling operating point. In this example, the initial OOV modeling operating point is fixed at a correct acceptance rate of 99.2% with a false acceptance rate of 71%. From this point we can then generate the remainder of the ROC curve by adjusting the confidence score accept/reject threshold. The solid line shows the optimal ROC curve for the combined approach which is generated by sampling results from all combinations of all operating points for the two different methods and extracting the "envelope" of these dual operating points. These curves demonstrate the significant improvement that can be obtained by the combined approach.

To further illustrate the improvement that can be obtained, suppose we wish to operate our system at a correct acceptance rate on keywords of 98%. At this operating point, the figure shows that the combined approach can reduce the false acceptance rate of misrecognized keywords by over 25% from either of the two original methods (from 55% to 40%). Although not shown, similar (though smaller) improvements can also be observed on the more general task of identifying recognition errors across all words. However, it is important to note that we focused on the keywords because they are most relevant to the overall understanding rate of the full system.

6. CONCLUSIONS & FUTURE WORK

This paper has presented a method for combining two techniques for detecting recognition errors. The first technique uses an explicit OOV model for detecting OOV words. The second approach relies on a confidence model to predict recognition errors. The combined approach uses OOV detection in a first stage and then confidence scoring in a second stage. In experiments comparing the two techniques, we found that the OOV modeling approach does better at detecting OOV words while the confidence scoring approach performs better in detecting misrecognitions in general. However, the combined approach shows significant improvement over either of the two approaches, especially for recognition of keywords.

There are numerous possible extensions to this work that we would like to examine in the future. One extension is to develop confidence methods specifically for determining whether a hypothesized OOV word is indeed OOV. A comparison of recognition paths containing and not containing a hypothesized OOV word could provide suitable confidence measures for making this decision. This would allow us to build a confidence model specifically for OOV word detection. A second possible extension is to examine different methods for combining the two approaches. Running parallel recognizers and using a post-processing voting scheme (i.e., ROVER [2]) is one possible alternative.

7. REFERENCES

[1] I. Bazzi and J. Glass, "Modeling out-of-vocabulary words for robust speech recognition," *Proc. of ICSLP*, Beijing, 2000.

[2] J. Fiscus. "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, 1997.

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

[4] A. Halberstadt and J. Glass, "Heterogeneous measurements and multiple classifiers for speech recognition," *Proc. of ICSLP*, Sydney, 1998.

[5] T. Hazen, *et al*, "Recognition confidence scoring for use in speech understanding systems," *Proc. of ISCA ASR2000 Tutorial and Research Workshop*, Paris, 2000.

[6] T. Hazen, *et al*, "Integrating recognition confidence scoring with language understanding and dialogue modeling" *Proc. of ICSLP*, Beijing, 2000.

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

[8] K. Kirchhoff and J. Bilmes, "Combination and joint training of acoustic classifiers for speech recognition," *Proc. of ISCA ASR2000 Tutorial and Research Workshop*, Paris, 2000.

[9] J. Macias-Guarasa, *et al*, "Acoustical and lexical based confidence measures for a very large vocabulary telephone speech hypothesis-verification system" *Proc. of ICSLP*, Beijing, 2000.

[10] A. Manos and V. Zue, "A segment-based spotter using phonetic filler models," *Proc. of ICASSP*, Munich, 1997.

[11] R. Rose and D. Paul, "A hidden Markov model based keyword recognition system," *Proc. of ICASSP*, Albuquerque, 1990.

[12] A. Wendemuth, G. Rose, and J. Dolfing, "Advances in confidence measures for large vocabulary," *Proc. of ICASSP*, Phoenix, 1999.

[13] 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.