ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

MODELING OUT-OF-VOCABULARY WORDS FOR ROBUST SPEECH RECOGNITION[1]

Issam Bazzi and James R. Glass

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139, USA
{issam,jrg}@sls.lcs.mit.edu

ABSTRACT

In this paper we present an approach for modeling and recognizing out-of-vocabulary (OOV) words in a single stage recognizer. A word-based recognizer is augmented with an extra OOV word model, which enables the OOV word to be predicted by a word-based language model. The OOV model itself is phone-based, so that an OOV word can be realized as an arbitrary sequence of phones. A phone bigram is used to provide phonotactic constraints within the OOV model. A recognizer with this configuration can recognize words in the original vocabulary as well as any potential new words of arbitrary pronunciation. In our preliminary investigation of this framework, we have evaluated the recognizer on a weather information domain with one test set containing only in-vocabulary (IV) data, and another containing OOV words. On the IV test set, the recognizer had an OOV insertion rate of only 1.3%, and degraded the baseline WER from 10.4% to 10.7%. On the OOV test set, the recognizer was able to detect nearly half of the OOV words (47% detection rate).

1. INTRODUCTION

Out-of-vocabulary (OOV) words are a common occurrence in many speech recognition applications, and are a known source of recognition errors [2]. For example, in our JUPITER weather information domain the OOV rate is approximately 2%, and over 13% of the utterances contain OOV words [12]. JUPITER utterances containing OOV words have a word error rate (WER) of 51%, while those containing only in-vocabulary words have a much lower WER of 10.4%. Although part of the increased WER on these OOV data is due to out-of-domain queries and spontaneous speech artifacts such as partial words, it is true that OOV words contribute to the increased WER. Since recognition errors are a typical source of mis-understanding, it is clearly important to improve the performance of the speech recognizer on OOV utterances. The ability to detect the location of OOV words would help significantly. In the past, we have used both sentence- and word-level confidence scoring to identify problematic utterances, such as those containing OOV words [10, 5]. In this work, we consider another tactic by incorporating an explicit OOV word model as part of the recognizer itself.

In addition to using the OOV model to detect the presence of OOV words, it is highly desirable to accurately recognize the sub-word units of the OOV word itself, so that a secondary analysis might actually hypothesize the OOV word. In this regard this work is a continuation of our efforts to develop a two-stage recognizer which has a domain-independent first-stage and is capable of processing arbitrary word sequences into a set of words or sub-word units for subsequent analysis by a domain-dependent second-stage recognizer. Such an architecture would allow many different spoken dialogue systems to share the same first-stage recognizer. The obvious challenge for such a configuration is to incorporate as much domain-independent constraint into the first stage as possible, so as to minimize any degradation in performance which will arise due to the lack of domain-specific constraints in the first stage.

In our preliminary investigations we considered the use of homogeneous sub-word lexical units (phones and syllables) in the first-stage recognizer [1], and found that syllables performed almost as well as words when provided domain-dependent data for training. In this work we are exploring the use of a hybrid approach which allows both word and sub-word units to exist in the first-stage. In this approach the recognizer combines word and sub-word units by building a model of a *generic* word in terms of the sub-word units. Since the sub-word units are a closed set covering all possible word sequences, the addition of words to the first-stage recognizer serves to provide additional constraint via a word-level language model.

In this paper we do not test the domain-independent aspect of the first stage. Instead we compare the performance of the hybrid recognizer configuration to a baseline word recognizer for a specific domain to measure the degradation in performance on in-vocabulary data, and evaluate its behavior on data containing OOV words. In the remainder of the paper we first provide an overview on related work. We then describe details of the system architecture, the generic word model, and the hybrid system. Finally, we present and discuss the results of a set of experiments in the JUPITER domain.

2. THE OOV PROBLEM

There are three different problems which can be associated with OOV words. The first problem is that of detecting the presence of an OOV word(s). Given an utterance, we want to find out if it has any words that the recognizer does not have in its vocabulary. The second problem is the accurate recognition of the underlying sequence of sub-word units (e.g., phones) corresponding to the OOV word. The third problem is the sound-to-letter problem, which might involve converting the sub-word sequence into an actual word so that it may be understood semantically [9].

[1]This material is based upon work supported by the National Science Foundation under Grant No. IRI-9618731.

Most of the work in the literature addresses the first problem, that is the detection of OOV words. The most common approach is to incorporate some form of filler or garbage model which is used to absorb OOV words and non-speech artifacts. This approach has been effectively used in key-word spotting for example, where the recognizer vocabulary primarily contains key-words, so that the filler models are used extensively [11, 8]. In these applications, non key-words absorbed by the filler model are of little subsequent interest. Our work differs from these applications in that we are very interested with accurately recovering the underlying sub-word sequence of an OOV word for the purpose of ultimately recognizing the word. Although in this paper we start with a simple phone-based model, and do not evaluate its accuracy, we are ultimately interested in increasing the complexity of the OOV model by incorporating additional sub-word structure, so that we can accurately recognize OOV words while not degrading the performance of the word-based recognizer.

3. MODELING OOV WORDS

In this section, we give an overview of the baseline word recognizer, the generic model used for OOV words, and the hybrid recognizer that combines a word system with the generic model to allow for OOV words in a single-stage recognizer.

3.1. The Baseline Word Recognizer

The word-based recognizer is based on the SUMMIT segment-based speech recognition system [4]. Typical recognizer configurations deploy a bigram language model in a forward Viterbi search, while a trigram (or higher-order) language model is used in a backward A^* search. The SUMMIT system uses a weighted finite-state transducer (FST) representation of the search space in which recognition can be viewed as finding the best path(s) in the composition

$$S = P \circ L \circ G, \tag{1}$$

where P represents the scored phonetic graph, L is the lexicon mapping pronunciations to lexical units, and G is the language model. The basic topology of the recognizer, illustrated in Figure 1, implies that traversing the network requires going through one or more words in the vocabulary. This is represented with words $w_1 \ldots w_n$ for the vocabulary and the loop back transition allowing for an arbitrary number of words.

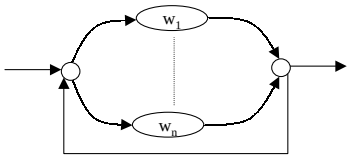

Figure 1: Search network for the word recognizer.

3.2. The Generic Word Model

Since an OOV word can consist of any sequence of phones (subject to language constraints), the generic word must allow for arbitrary phone sequences during recognition. One of the simplest word models is a phone recognizer; one whose vocabulary is made of the set of phones for the language. Since this unit inventory can cover all possible words, it can be used as the basis for the generic word. The phone inventory also has the advantage of being small in size.

In FST terms, a phone recognizer can be represented as:

$$S = P \circ L_p \circ G_p \tag{2}$$

where L_p and G_p are the phone lexicon and grammar, respectively. For our phone recognizer, L_p is a trivial FST and can be discarded, since the phone units in P are already the basic units of the word lexicon. The phone grammar, G_p, can consist of a phone-level n-gram language model. Figure 2 shows the network corresponding to such a configuration. Similar to Figure 1, the search network allows for any sequence of phones consisting of $P_1 \ldots P_n$. The generic word model based on a phone

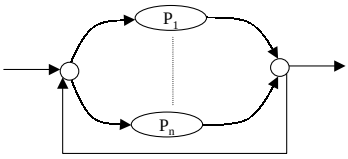

Figure 2: A generic word model based on phones.

recognizer is constrained only by the phone grammar that biases different paths in the network. The phone level language model used here is trained on phone sequences from the training corpus where words are replaced with their phonetic pronunciation. One consequence of this approach is that certain bigram pairs could be cross-word pairs, that is the first phone is the end of one word and the second phone is the start of another. A variation on this approach would be to train the n-gram on sequences of phones of individual words, making the grammar more tuned to within-word phone sequences rather than cross-word sequences.

There are several ways to incorporate additional constraints or structure into the generic word model. One way is to use larger sub-word units such as syllables or morphs [3]. Syllables will increase the size of the generic word model because there are many more syllables than phones, but, as we observed previously [1], they provide more constraint for decoding OOV words provided all syllables in the language are included. Another way to incorporate more structure on the phone topology of the generic word is to impose a minimum duration requirement on the size of the word. For this paper however, we explore only the phone recognizer as the generic word model for OOV detection and recognition. In the following subsection we show how to integrate the generic word model with the baseline word recognizer to allow for OOV words during the search.

3.3. The Hybrid Configuration

To create the hybrid recognizer we add to the baseline word recognizer's vocabulary an OOV word whose underlying model is the generic word model presented previously. Figure 3 shows how the word search space can be augmented with the generic word model. We simply allow the search to have a transition to enter into the generic word model W_{OOV}. As we exit W_{OOV}, we are allowed to either end the utterance or enter into any other word, including the OOV word. The transition into the generic

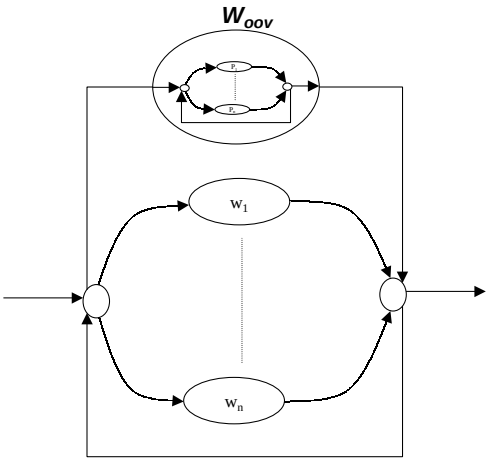

Figure 3: The hybrid recognition configuration.

word model can be controlled via an OOV penalty (or cost) C_{OOV}. This penalty 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 language model of the hybrid recognizer remains word-based, but must now include an OOV entry for unknown words. Since the OOV word is part of the vocabulary, the grammar will include n-grams with OOV words that will be used during the search just like transitions into any other word in the vocabulary.

As mentioned in Section 2, augmenting the word recognizer with the generic word model as shown in Figure 3 is somewhat similar to using filler (or garbage) models for word-spotting. 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. In most word spotters however, that use a filler model, the effective vocabulary is much smaller, so that most input words are covered by the filler model. The second distinction is that accurate sub-word recognition is important for our OOV model since we intend to use its output for a second stage of processing to identify the OOV word. In contrast, word spotters typically make no use of the output of the filler models.

The hybrid recognizer can be represented with FSTs as follows:

$$S_H = P \circ (L \cup (L_p \circ G_p \circ T_{OOV}))^* \circ G' \qquad (3)$$

where S_H is the hybrid search space. T_{OOV} is the topology of the OOV word which for the phone recognizer is a single state FST with a self loop allowing for any arbitrary sequence of phones. G' is simply the same as G except for the extra unknown word in the vocabulary. That is when an unknown word is encountered in the training of G', the word is considered to be W_{OOV} and gets treated like any other word in the vocabulary, so the n-gram will have bigram pairs such as (W_m, W_{OOV}) and (W_{OOV}, W_n) for some words W_m and W_n.

The search space S_H relies mainly on the union of the two search spaces. The union operation, $\cup$, provides the choice of going through either the word network from the vocabulary or through the generic word network. The $*$ operation is the closure operation on FSTs. This operation allows for switching between the two networks during the same search allowing for the transition into and out of the OOV network as many times as needed.

C_{OOV}	OOV Detection (%)	IV False Alarm (%)	IV WER (%)
∞ (Baseline)	0	0	10.4
0	46.8	1.3	10.7
-1	54.4	3.2	10.8

Table 1: Detection and false alarm for $C_{OOV} = 0, -1$.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

All the experiments for this work are within the JUPITER weather information domain [12]. The baseline system used a similar configuration to that which has been reported previously [4]. A set of context-dependent diphone acoustic models were used, 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. Bigram language models were used both at the word-level, as well as at the phone-level for 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. There were two test sets used to evaluate the recognizers. The first test set consisted of a set of 400 utterances containing only in-vocabulary (IV) words. The second test set consisted of 314 utterances which contained at least one OOV word (most of the OOV utterances had only one OOV word).

4.2. Results

We ran a series of experiments on the two test sets described above. Our main goal was to demonstrate whether this approach can detect OOV words without significantly degrading the performance of the word recognizer on IV utterances. For this reason we measured word error rates (WERs) and OOV false detection (alarm) rates on the IV data, although these two measures are correlated. We also measured the OOV detection rate on the OOV test data to see how well we could detect OOV words.

Detection of an OOV word is assumed when the top hypothesis of the recognition chooses a path through the generic word model. Absence of an OOV word from the best hypothesis indicates that no OOV word was detected. There are other ways to define OOV detection by looking at the frequency of the OOV word in the N-Best as opposed to only the best hypothesis. We experimented only with the first approach.

For the series of experiments we present here, we varied the OOV penalty C_{OOV}. Table 1 shows the results for two values of C_{OOV} (0 and -1). The second column shows the OOV detection rate on the OOV test set, the third column shows the false

alarm rate on the IV test set and the fourth column shows the IV word error rate (WER). As the table shows, with no OOV penalty we detect nearly half of the OOV words while have a small amount of false OOV detections on the IV data. We were also pleased to observe a very small degradation in overall WER from the baseline word recognizer from 10.4% to 10.7%.

Figures 4 shows the Receiver Operating Characteristics (ROC) curve for several values of C_{OOV} ($-\infty$,-5,1,2,3,4,5,$+\infty$). Figure 5 shows the WER for the IV test set as the false alarm rate increases on the IV data. As expected performance significantly degrades for high false alarm rates.

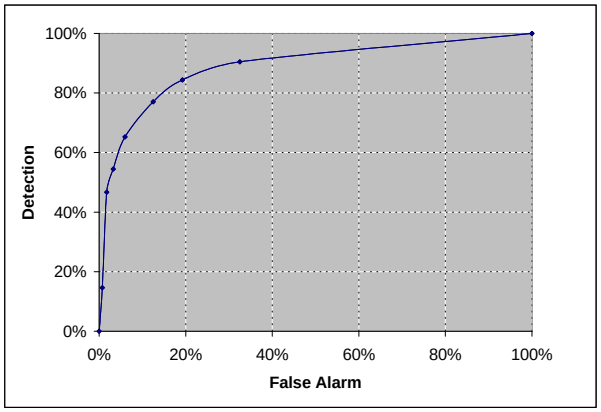

Figure 4: ROC curve for OOV detection.

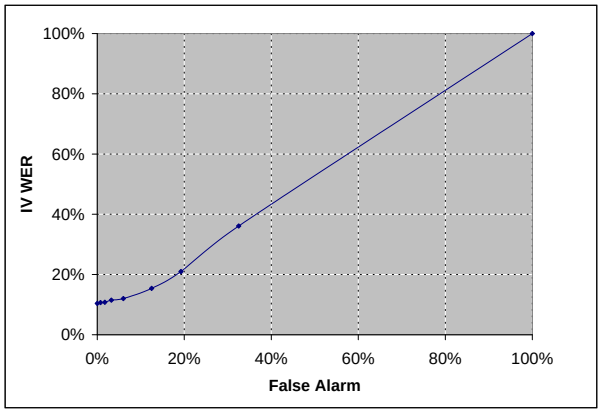

Figure 5: IV WER versus false alarm rate.

5. DISCUSSION AND FUTURE WORK

The results we obtained so far are quite encouraging. With a very simple generic word model, we were able to detect half of the OOV words with a very small degradation in WER as well as a low false alarm.

Augmenting the word based lexicon with the generic word model should theoretically increase the size of the FST models by that of the phone recognizer. For the experiments we presented the size of the models increased by roughly a factor of three even though the phone recognizer is only a fraction of the size of the word recognizer. We attribute this significant increase in size to the way the phone FST was augmented and then optimized (replicas of the phone recognizer could have been created during optimization). We continue to work on the augmentation procedure to ensure only a small increase in the final model size. This will be essential when augmenting large generic word models such as a syllable recognizer.

For our current work, we are working on incorporating a probabilistic duration model for OOV words. This duration model will require a minimum number of phones for an OOV word as well as probability scores for different word lengths. Another aspect of the approach we are working on is the use of larger units (syllables) to model the OOV word. Syllables should provide a more robust sub-word unit to model generic words. In addition, we are considering the use of classes of OOV words (instead of only one) such as an OOV model for city names, another for weather terms, and so on.

In related work [5], a word-level confidence measure is used to detect mis-recognized words, among which (of-course) are the OOV words. We plan to investigate combining the results of our hybrid recognizer with confidence measures to achieve better OOV detection. Finally, we will be looking at the last part of the OOV problem that of proposing real words or semantic properties to recognized OOV words based on their phone sequence.

Acknowledgments Lee Hetherington and T.J. Hazen provided many suggestions, and helped implement tools for this research.

6. REFERENCES

1. I. Bazzi and J. Glass, "Heterogeneous Lexical Units for Automatic Speech Recognition: Preliminary Investigations," *Proc. ICASSP*, Istanbul, 1257–1260, 2000.
2. L. Chase, "Error-Responsive Feedback Mechanisms for Speech Recognizers," Ph.D. Thesis, Carnegie Melon University, 1997.
3. G. Chung and S. Seneff, "Improvements in speech understanding accuracy through the integration of hierarchical linguistic, prosodic, and phonological constraints in the JUPITER domain," *Proc. ICSLP*, Sydney, 935–938, 1998.
4. J. Glass, J. Chang, and M. McCandless, "A probabilistic framework for feature-based speech recognition," *Proc. ICSLP*, Philadelphia, 2277–2280, 1996.
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. D. Klakow, et al., "OOV-Detection in Large Vocabulary System Using Automatically Defined Word-Fragments as Fillers," *Proc. Eurospeech*, Budapest, 49–52, 1999.
7. R. Lacouture and Y. Normandin., "Detection of ambiguous portions of signal corresponding to OOV words or misrecognized portions of input," in *Proc. ICSLP*, Philadelphia, 2071–2074, 1996.
8. A. Manos and V. Zue, "A Segment-based Spotter Using Phonetic Filler Models," *Proc. ICASSP*, Munich, 899–902, 1997.
9. H. Meng, "Phonological Parsing for Bi-directional Letter-to-Sound/Sound-to-Letter Generation," Ph.D. Thesis, MIT, 1995.
10. C. Pao, P. Schmid, and J. Glass, "Confidence scoring for speech understanding," *Proc. ICSLP*, 815–818, Sydney, 1998.
11. R. Rose and D. Paul, "A Hidden Markov Model Based Keyword Recognition System," *Proc. ICASSP*, Albuquerque, 129–132, 1990.
12. V. Zue, et al., "JUPITER: A telephone-based conversational interface for weather information," *Proc. SAP*, 88(1), 85–96, Jan., 2000.