

HYBRID LANGUAGE MODELS FOR OUT OF VOCABULARY WORD DETECTION IN LARGE VOCABULARY CONVERSATIONAL SPEECH RECOGNITION

*Ali Yazgan**

Center for Language and Speech Processing
Johns Hopkins University
3400 N. Charles St. Baltimore, MD 21218
ayazgan1@jhu.edu

Murat Saraclar

AT&T Labs – Research
180 Park Avenue
Florham Park, NJ 07932-0971, USA
murat@research.att.com

ABSTRACT

In this paper, we propose a method for out-of-vocabulary (OOV) word detection and taking a step toward open vocabulary automatic speech recognition. The proposed method uses a hybrid language model combining words and sub-word units such as phones or syllables. We describe a detection algorithm based on the posterior count of the OOV words given the hybrid model, and compare it to using the posterior probability of the best word string given a conventional word only model. Experimental results on the Switchboard corpus are presented for different vocabulary sizes. The new method yields a gain of over 10% in OOV word detection. In addition, a modest number of the OOV word pronunciations are found correctly.

1. INTRODUCTION

Almost all automatic speech recognition (ASR) systems have a closed vocabulary. This restriction comes from run-time requirements as well as the finite amount of data used for training the language models of the ASR systems. Typically the recognition vocabulary is taken to be the words appearing in the language model training corpus. Sometimes the vocabulary is further reduced to only include the most frequent words in the corpus. The words that are not in this closed vocabulary – the out of vocabulary (OOV) words – will not be recognized by the ASR system, contributing to recognition errors.

Although OOV words are known to be a major source of recognition errors, the problem is simply ignored in most large vocabulary ASR systems. This is commonly justified by citing low OOV rates by token. For tasks where transcription is the end goal and word error rate (WER) is the performance measure this justification is reasonable. However, for automatic indexing, searching and browsing of spoken communications, the ASR system acts like a front

end to convert audio into text, and the end goal is retrieval. The common performance metrics for retrieval are based on precision and recall, and this makes OOV rate by type more relevant. As will be shown here this rate can be quite high for conversational speech, making the OOV problem an important one. The effects of OOV words in spoken document retrieval is discussed in [1]. Similar arguments can be made for speech understanding tasks where named entity extraction is vital. In addition, OOV words are often semantically rich, which makes OOV detection desirable.

There is some recent work done on this topic. The basic idea of most is representing the OOV words with sub-word units. For example, the work presented in [2] uses a separate phone language model for the OOV words which is merged with the in-vocabulary word language model. A combination of this model with confidence scoring is given in [3]. An extension using automatically learned multi-phone units for use within the separate OOV model is presented in [4]. Further detail about this approach for modeling OOV words and experimental results on both JUPITER weather information domain and Broadcast News (Hub4) can be found in [5]. A very similar approach to ours is given in [6] which uses pairs of graphemes and phonemes as sub-word units. Reported experimental results are on Broadcast News (Hub4). Both of these approaches use only the ASR 1-best output. There is also some work done in order to improve readability of text by transcribing OOV words based on phoneme-to-grapheme conversion [7].

In this paper we investigate using hybrid language models for detection of utterances containing OOV words. We introduce a detection algorithm based on ASR lattices, as well as 1-best hypotheses. We also aim to find pronunciations for detected OOV words. Unlike prior work, we report results on conversational speech (Switchboard) which is more challenging. In section 2 we explain our approach to the OOV word detection problem. Next, details of our experimental setup is given in section 3. Results are presented in section 4. We make our conclusions in Section 5.

*The research reported in this paper was performed while the author was at AT&T Labs – Research.

2. METHOD

2.1. The Baseline ASR System

The ASR system used in this work is based on AT&T's Switchboard Evaluation system [8]. The language model is estimated from the Switchboard training corpus using the AT&T GRM Library. In order to understand the corpus characteristics with respect to OOV words and to investigate the effects of OOV words on recognition performance, we build language models for various vocabulary sizes. The words that appear in the training corpus but are not in the recognition vocabulary are mapped to a unique OOV token. Although the OOV token is used during estimation it is ignored during recognition.

2.2. Hybrid Language Models

Our solution to the OOV detection problem uses a hybrid language model combining words and sub-word units. In this study, the sub-word units are taken to be either phones or syllables. In our first model we combine words with phones and in the second one we combine words with syllables. Instead of estimating separate word and sub-word language models which are subsequently combined, we estimate a single language model containing both words and sub-word units. First we partition the recognition vocabulary into two subsets. The first set contains the most frequent N words and the second set contains the rest of the recognition vocabulary. Each word in the training corpus that belongs to the second set is mapped to its pronunciation(s) in terms of the sub-word units. This transformed corpus is then used for estimating the hybrid language model. As an example, if the word OUTFIT is in the second set, then the sentence

WHAT TYPE OF OUTFIT DO YOU HAVE ...

will be transformed to

WHAT TYPE OF /aw/ /T/ /f/ /ih/ /T/ DO YOU HAVE ...

for the word and phone hybrid model, and to

WHAT TYPE OF aw_T/ f_ih_T/ DO YOU HAVE ...

for the word and syllable hybrid model. The syllabification is done automatically using the maximum onset principle.

Of course, the pronunciation dictionaries need to be extended to account for the transformation. We add all the phones and syllables to the pronunciation dictionary with the appropriate pronunciations. Example entries in the final dictionary are:

OUTFIT \rightarrow /aw/ /T/ /f/ /ih/ /t/

/aw/ \rightarrow /aw/

aw_T/ \rightarrow /aw/ /T/

This approach has some advantages over methods based on combination of separately trained word and sub-word models. First, the sub-word portion of the language model

is estimated only from less frequent words, which provide a better match for OOV words. Second, dependences between word and sub-word units are better captured since there is no forced back-off during the transition points. Finally, there is no need for a scaling factor or insertion penalty while combining the models.

2.3. OOV Detection

We investigate OOV detection using lattices as well as the 1-Best hypothesis generated with the hybrid language models. In the 1-Best case, an utterance that contains a sub-word string after a filtering step is declared to contain an OOV word. In the filtering step we eliminate all sub-word strings exactly corresponding to a word in the pronunciation dictionary. We also discard all "short" phone strings (less than three phones).

In the lattice case, after the filtering step we map the remaining phone strings to a special OOV symbol. For each path x in the lattice L , we define the path OOV count $C(OOV|x, L)$ to be the number of times the OOV symbol is seen on that path. The lattice OOV count $C(OOV|L)$ is defined as

$$C(OOV|L) = \sum_{x \in L} p(x|L) C(OOV|x, L)$$

where $p(x|L)$ is the posterior probability of path x given the lattice L . Our hypothesis is that the utterances with high $C(OOV|L)$ will have a high probability of containing an OOV word. Then, the detection is done by comparing $C(OOV|L)$ to a threshold t .

As a baseline system, we use a detection approach based on utterance posterior probabilities using a conventional word-only language model. Our hypothesis is that utterances with low posterior probabilities in the ASR output usually have errors. Some of these errors are caused by OOV words. We find the posterior probability of the best path in the ASR lattice output and decide whether it contains an OOV word or not. In other words, we use utterance level confidence scores to detect utterances with OOV words.

3. EXPERIMENTAL SETUP

All the experiments are done with the RT02 Switchboard Evaluation data with 67066 word tokens and 3820 word types and a total of 6266 utterances. Statistics for different vocabulary sizes is given in Table 1.

Table 2 gives information about the experiment sets A, B and C. For example, in experiment A, the IV size (or N) is 2K and the test data vocabulary is 5K. Every word in frequency range 2000-5000 is mapped to its sub-word units string. Each word outside the 5000 word set is treated as an OOV word.

Vocabulary Size	OOV by type %	OOV by token %	Utterances with OOV %	WER %
5K	32.5	3.0	19.4	41.5
10K	18.7	1.7	11.2	40.4
20K	10.4	1.0	6.4	40.1
45K	6.2	0.7	3.6	40.1

Table 1. Switchboard test data statistics

Exp	IV (N)	Words	Sub-Word Units	Test data Vocabulary
A	2K	Between 0-2K	Between 2K-5K	5K
B	5K	Between 0-5K	Between 5K-10K	10K
C	5K	Between 0-5K	Between 5K-20K	20K

Table 2. Details for experiment sets. IV is the in-vocabulary size (N).

4. RESULTS

In this study we use two pairs of performance measures for evaluation. The first pair is OOV detection and false alarm, and the second pair is precision and recall. OOV detection rate (Det) is defined as the number of utterances detected correctly as having OOV words divided by the number of utterances with OOV words. False alarm rate (FA) is the number of utterances detected incorrectly as having OOV divided by the number of utterances without OOV words. A good model has a low false alarm rate and a high OOV detection rate.

Recall is the same as OOV detection rate. Precision is defined as the number of utterances detected correctly as having OOV words divided by total utterances detected as having OOV. A good model has high precision and recall.

4.1. OOV Word Detection Using 1-Best Hypotheses

First, in Table 3, we report results with a detection system using the ASR 1-best output for the three experiment sets with different vocabulary sizes and sub-word unit types.

As seen in Table 3, for experiment sets B and C (where $N=5K$) we achieve a 41.5% WER with the word-phone hybrid model. Recall from Table 1 that the WER for the 5K word model is also 41.5%, indicating that the hybrid system models the OOV word without deteriorating the WER performance of the word-only model.

Experiment	Sub-word Units	WER (%)	FA (%)	Det (%)
A	phones	43.2	17.9	75.7
B	phones	41.5	13.5	58.8
C	phones	41.5	20.6	69.1
A	syllables	45.0	16.4	65.0
B	syllables	43.9	10.7	47.2
C	syllables	43.6	16.1	59.4

Table 3. OOV Detection Using Best Path. Word Error Rate (WER), False Alarm (FA) vs. OOV Detection rate (Det) for different experiments A, B and C.

4.2. OOV Word Detection Using Lattices

For lattice based OOV word detection, we obtain different operating points by varying the threshold used in detection. The performance at these operating points are plotted as curves. We present results for the baseline (“word”) detection system, the word-phone hybrid system (“hybrid-phns”) and the word-syllable hybrid system (“hybrid-syls”). Baseline curves are obtained by comparing the posterior probability of each utterance to changing threshold values. The hybrid model curves are obtained by comparing the OOV posterior count of each utterance to changing threshold values.

Figure 1 shows the false alarm vs. OOV detection performance of the systems on experiment set A. The hybrid model with words and phones performs 10-15% better over the baseline. The hybrid model combining words and syllables performs slightly worse than this model. Figure 2 shows the precision vs. recall performance of the systems on the same test set. We observe a significant improvement over the baseline (i.e 10-20%). The word-phone hybrid model again performs better than the word-syllable hybrid model. We get similar curves for experiment sets B and C (not shown due to space limitations).

Since it uses a tunable threshold, the ASR lattice based detection system is more flexible than the system using only ASR 1-best. In addition, using lattices improves the OOV detection rate by 1-3% for the same false alarm rate.

4.3. Finding the Pronunciation

After detecting a good portion of OOV words with low false alarm rates our goal is to determine the pronunciations of these words. With our OOV detection system using ASR 1-best with word-phone hybrid models, we align the reference word transcription with the ASR output. We then compare the aligned pronunciations with the ones in the pronunciation dictionary. It turns out that out of all detected OOV words in experiment set A, we can find 7.5% of the pronunciations exactly. Some of the correct pronunciations we obtain are given in Table 4.

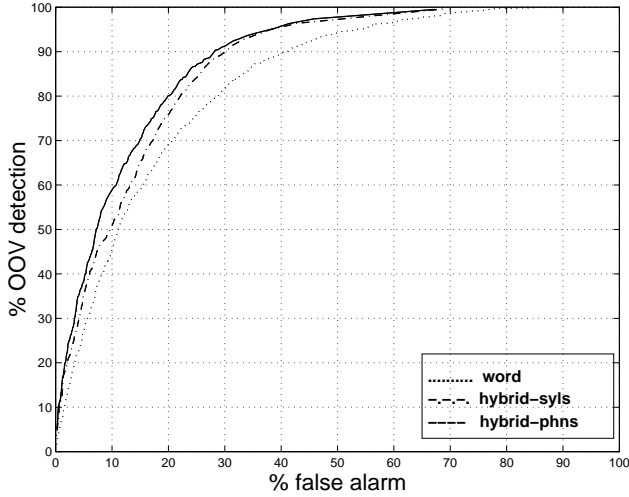


Fig. 1. False alarm vs. OOV detection rate on 5K test data

Word	Detected Pronunciation
LEGALIZING	/l/ /iy/ /G/ /ax/ /l/ /ay/ /z/ /ih/ /ng/
CYCLICAL	/s/ /ih/ /K/ /l/ /ih/ /K/ /ax/ /l/

Table 4. Examples of correct pronunciations.

An additional 7.5% of OOV pronunciations have only a single phone error. Some of the most common error patterns are presented in Table 5. For the first example, the error is a minor vowel substitution. In the second and third examples there is an additional phone at the beginning or at the end. Sometimes frequent short words cause phone deletions, as seen in the last example.

Word	Detected Pronunciation
HIGHLANDER	/hh/ /ay/ /l/ /ae/ /n/ /D/ /er/
DISTRACTION	/ax/ /D/ /ih/ /s/ /T/ /r/ /ae/ /K/ /sh/ /ih/ /n/
INTEL	/ih/ /n/ /T/ /eh/ /l/ /iy/
HEALER	HE /ax/ /l/ /er/

Table 5. Examples of pronunciations with errors.

5. CONCLUSION

In this paper, we presented a method for OOV word detection and finding OOV word pronunciations in a given utterance. The method uses a hybrid language model (either words with phones or words with syllables) and has a decision criterion based on the posterior OOV count for every utterance. As a baseline system we use conventional word-only language models and the posterior probability of each utterance. 10% to 15% improvement in OOV detection is obtained over the baseline for a significant range of false alarm values. Similarly a 10-20% improvement in precision is observed for a wide range of recall values. Using lattices

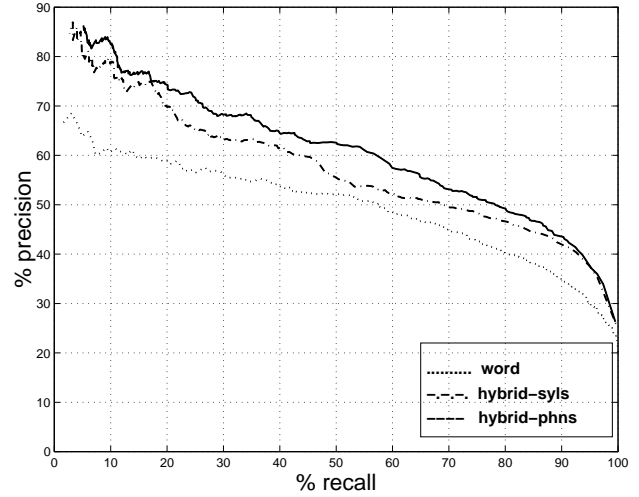


Fig. 2. Precision vs. Recall curve on 5K test data

yields an improvement of 1-3% over using the best hypotheses. Out of all detected OOV words we can find 7.5% of the pronunciations just by looking at the best path.

6. REFERENCES

- [1] P.C. Woodland, S.E. Johnson, P. Jourlin and K.Sparck Jones, "Effects of Out of Vocabulary Words in Spoken Document Retrieval," *Proc. SIGIR*, pp 372-374, Athens, Greece, 2000.
- [2] I. Bazzi and J. Glass, "Modeling Out of Vocabulary Words for Robust Speech Recognition," *Proc. of ICSLP* Beijing, 2000.
- [3] T.J. Hazen and I. Bazzi, "A comparison and Combination of methods for OOV Word Detection and Word Confidence Scoring," *Proc. of ICASSP* Salt Lake City, 2001.
- [4] I. Bazzi and J. Glass, "Learning Units for Domain-Independent Out-of-Vocabulary Word Modeling," *Proc. of Eurospeech* Aalborg, 2001.
- [5] I. Bazzi, "Modelling Out-of-Vocabulary Words for Robust Speech Recognition," *PhD Thesis* MIT, 2002.
- [6] L. Galescu, "Recognition of Out of Vocabulary Words with Sub-Lexical Language Models," *Proc. of Eurospeech* Geneva, Switzerland, 2003.
- [7] B. Decadt, J. Duchateau, W. Daelemans, P. Wambacq, "Transcription of Out-Of-Vocabulary Words in Large Vocabulary Speech Recognition Based on Phoneme-To-Grapheme Conversion," *Proc. Of ICASSP*, vol 1, pp 861-864, Orlando, Florida, USA, 2002.
- [8] A. Ljolje et al. "AT&T Switchboard Evaluation System", *RT03 Workshop*, Boston, 2003.