Towards Using Hybrid Word and Fragment Units for Vocabulary Independent LVCSR Systems

Ariya Rastrow[1], *Abhinav Sethy*[2], *Bhuvana Ramabhadran*[2] *and Frederick Jelinek*[1]

[1]Human Language Technology Center of Excellence, and
Center for Language and Speech Processing, Johns Hopkins University
[2]IBM T.J. Watson Research Center, Yorktown Heights, NY, USA
{ariya, jelinek}@jhu.edu {asethy, bhuvana}@us.ibm.com

Abstract

This paper presents the advantages of augmenting a word-based system with sub-word units as a step towards building open vocabulary speech recognition systems. We show that a hybrid system which combines words and data-driven, variable length sub word units has a better phone accuracy than word only systems. In addition the hybrid system is better in detecting Out-Of-Vocabulary (OOV) terms and representing them phonetically. Results are presented on the RT-04 broadcast news and MIT Lecture data sets. An FSM-based approach to recover OOV words from the hybrid lattices is also presented. At an OOV rate of 2.5% on RT-04 we observed a 8% relative improvement in phone error rate (PER), 7.3% relative improvement in oracle PER and 7% relative improvement in WER after recovering the OOV terms. A significant reduction of 33% relative in PER is seen in the OOV regions.

Index Terms: hybrid LVCSR system, out-of-vocabulary, subword unit selection, phone recognition, OOV detection

1. Introduction

Many technologies used in Spoken-Term Detection (STD) and retrieval, out-of-vocabulary (OOV) detection, universal phone recognition and speaker and language identification, require accurate recognition of phonemes. In indexing and retrieval of spoken documents, OOV queries are typically detected using various phonetic recognition based approaches[1, 2, 3, 4, 5]. A better phone accuracy in the in-vocabulary (IV) regions and OOV regions would improve the overall term detection accuracy by reducing both false detections and misses.

Several techniques have been proposed in acoustic modeling literature for improving phone recognition accuracy [6]. An alternate solution for improving phone recognition accuracy is to increase the lexicon size in a large-vocabulary continuous speech recognition (LVCSR) system. However, blindly enlarging the dictionary size, can result in an increased Phone error rate (PER) and Word error rate (WER), as a result of increased confusability amongst the newly added words. In this paper, we use a hybrid recognition system which combines words and sub-word units [7] (instead of adding new words) to improve the phone recognition accuracy. Sub-word units unlike words, are not tied to a specific domain/genre. It is expected that the hybrid system can be easily extended to any domain of a given language and the sub-word units themselves can be trained on any generic domain where sufficient training material is available. We also show that by using these sub word units along with the words, one can detect OOV regions and recover the OOV word(s) from the output of a LVCSR system. Reliable detection of the presence and location of the OOV words can be used to improve the performance of real world applications of automatic speech recognition systems.

This paper focusses on detecting OOV regions and improving phone recognition accuracy. The rest of the paper is organized as follows. Section 2 describes the definition of the sub-word units and the derivation of a hybrid language model. Section 3 describes the construction of the hybrid and word based LVCSR systems with varying vocabulary sizes and the algorithm for the computation of phone error rates (PER). Section 4 describes the training and test data sets for the hybrid and word-based LVCSR systems. The PER and OOV detection results are presented in Section 5. A mechanism for recovering OOV words from the output of a hybrid system is presented in Section 6. The paper concludes with suggestions for further research in Section 7.

2. Hybrid word/sub-word units in an LVCSR system

Fragments are sub-word units with variable length phone sequences and are selected automatically using statistical methods[7]. A hybrid LVCSR system uses the same acoustic models as the word-based LVCSR system while the language model is built from text that is tokenized into words and sub-word units.

2.1. Sub-word/Fragment Selection

Fragment[1] selection methods can be classified into two categories, namely, knowledge-driven methods that incorporate linguistic knowledge and data-driven methods [8, 7] which maximize an objective function. In this paper, the approach suggested in [7] is used. The LM training text is converted into phones using a pronunciation dictionary. All OOVs are excluded from the training set. Using this data set, an N-gram phone LM is built and pruned using a relative-entropy based method [9]. This results in a set of fragments comprised of phone sequences of length 1 to N.

2.2. Hybrid Language Model

The hybrid LM captures the dependencies between word and sub-word units. The LM training data is obtained by converting OOV terms in the text to their fragment representation. Pronunciations for the OOV terms are obtained using grapheme to

[1] Sub-word units and fragments, are used interchangeably throughout the paper

$< s >$ THE BODY OF *ZIYAD HAMDI* WHO HAD BEEN SHOT WAS FOUND SOUTH OF THE CITY $< /s >$
$< s >$ THE BODY OF *Z_IY Y_AE_D HH_AE_M D_IY* WHO HAD BEEN SHOT WAS FOUND SOUTH OF THE CITY $< /s >$

Table 1: Tokenized Hybrid LM text

phone models [10]. A greedy search algorithm begins by assigning the longest possible matching fragment and iteratively uses the next longest fragment until the entire pronunciation of the OOV term has been represented by sub-word units. We also experimented with other techniques for tokenizing the LM text based on the degree of confusability of the fragments with the pronunciation of in-vocabulary words, i.e. selecting only those fragments that are less confusable with the words in the dictionary. These studies have been inconclusive and the rest of the paper uses fragments selected by the greedy algorithm. A hybrid LM is built on the tokenized text treating each sub-word unit as an individual token.

Table (1) illustrates an example of tokenized hybrid text obtained using greedy search algorithm for tokenizing the LM text into sub-word units and words where terms *ZIYAD* and *HAMDI* are OOV.

3. LVCSR Systems, PER and OOV Detection

To study the effectiveness of the hybrid system on phone recognition accuracy and OOV detection, hybrid and word systems for different vocabulary sizes were built. Lexicons ranging from $10K$ words to 84K were selected by sorting the words in the word-based LM by their unigram probabilities and selecting the top n words that yielded a specific vocabulary size. The set of fragments was selected as described in Section 2.1 by using a 5-gram phone LM for each vocabulary size. The size of this set was fixed at $20K$ for all systems. Therefore, the hybrid system includes $20K$ fragments, in addition to the words in its lexicon.

Phone Error Rate (PER) computation is done using the NIST scoring tool *sclite*. The phone sequence in the hypothesis is aligned with the reference phone sequence with equal costs for substitution, insertion and deletion of phones. The reference phone sequence is obtained by forced alignment of the audio stream to the reference transcripts. The oracle phone error rate is computed using a dynamic programming based string alignment algorithm which minimizes the Levenshtein distance function.

OOV detection results are based on the work presented in [11]. In this setup, the posterior probability of the subword units inside confusion networks[12] decoded by the hybrid LVCSR system was found to be a good indicator for the presence of OOV regions.

4. Experimental Setup

The LVCSR system is based on the 2007 IBM Speech transcription system for GALE Distillation Go/No-go Evaluation [13]. The acoustic models are discriminatively trained on speaker-adapted PLP features. These acoustic models are used across all the experiments presented in this paper. The LM training text (for all systems) consists of 335M words from the following *broadcast news* (BN) data sources [13]: 1996 CSR Hub4 Language Model data, EARS BN03 closed captions, GALE Phase 2 Distillation GNG Evaluation Supplemental Multilingual data, Hub4 acoustic model training transcripts, TDT4 closed captions, TDT4 newswire, and GALE Broadcast Conversations and GALE Broadcast News. All word- and hybrid-based language models are 4-gram LMs with Kneser-Ney smoothing. This LM training text is also used select fragments. The best LM (with the 14.1% WER) built from a 84K lexicon with an average of 1.08 pronunciation variants per word has $3.3M$ n-grams and a perplexity of 204 on the RT-04 Dev set. The LVCSR system has a WER of 14.1% with a 84K lexicon on the RT-04 Eval set.

We report Phone Error Rate (PER), Word Error Rate (WER) and OOV detection results on the RT-04 Broadcast News Evaluation set (45K words) as an in-domain test set and the MIT lectures data set [14] (176K words, 21 hours, 20 lectures given by two speakers) as an out-of-domain test set.

5. Results

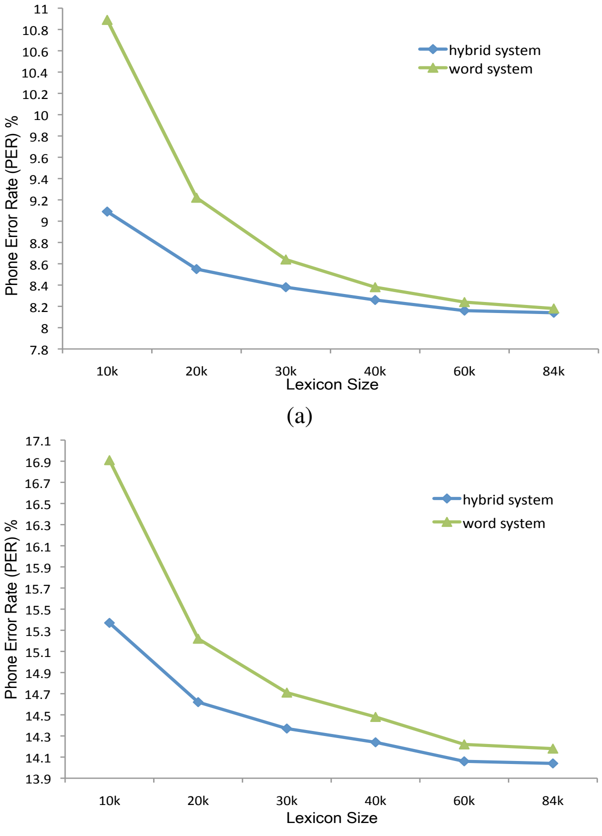
In order to study the value of adding fragments to word-only lexicons, LVCSR systems for various lexicon sizes were built. The lexicon was determined using the method described in Section 3. The OOV rates for these systems on the two test sets are reported in Table 2. Each of these lexicons was augmented with $20K$ fragments and hybrid systems were built for each system using the procedure described in 2.2.

Vocab. Size	10k	20k	30k	40k	60k	84k
RT-04 (%)	5.04	2.48	1.47	1.04	0.68	0.54
Lectures (%)	7.88	5.45	4.51	4.09	3.53	3.45

Table 2: OOV rates for the RT-04 set and the MIT lectures data

5.1. Phone Accuracy

Both word and hybrid systems were used to decode the two test sets described in Section 4. Figure 1 illustrates the PER results on the two test sets for both hybrid and word systems with different lexicon sizes. It can be seen that the hybrid system consistently outperforms the word system in phone recognition accuracy. The difference is more apparent in regions where the word system has a high OOV rate. For example, in the RT-04 test set, the $10K$ word system has a PER of 10.89%, while the hybrid system with the same $10K$ words and an additional set of fragments has a PER of 9.09%. Both systems converge to a PER of 8.1% with a full-blown lexicon of the best system ($84K$). Since the fragments were built on the broadcast news domain, we validated the robustness of this result on the MIT lecture data set. The $10K$ word system has a PER of 16.9% while the hybrid system with the additional fragments has a PER of 15.3%. The word system converges to a PER of 14.2% with a $84K$ lexicon while the hybrid system converges to a PER of 14.1%. Although the number of words in both the word and hybrid system is the same, the hybrid system includes $20K$ additional fragments and consequently has a larger decoding lexicon. Thus the above comparison is done on two systems of different lexicon sizes. The purpose of this comparison is merely to illustrate that a hybrid approach yields better phone recognition performance in new domains where only limited data is available to build exhaustive lexicons and train good language models. In such cases, one could use a smaller sized system built on commonly available data such as broadcast news to rapidly decode the data from the new domain, identify OOV regions (see section 5.2) and recover new entries for a lexicon that better fits the new domain. A first attempt at this is discussed in Section 6.



(a)

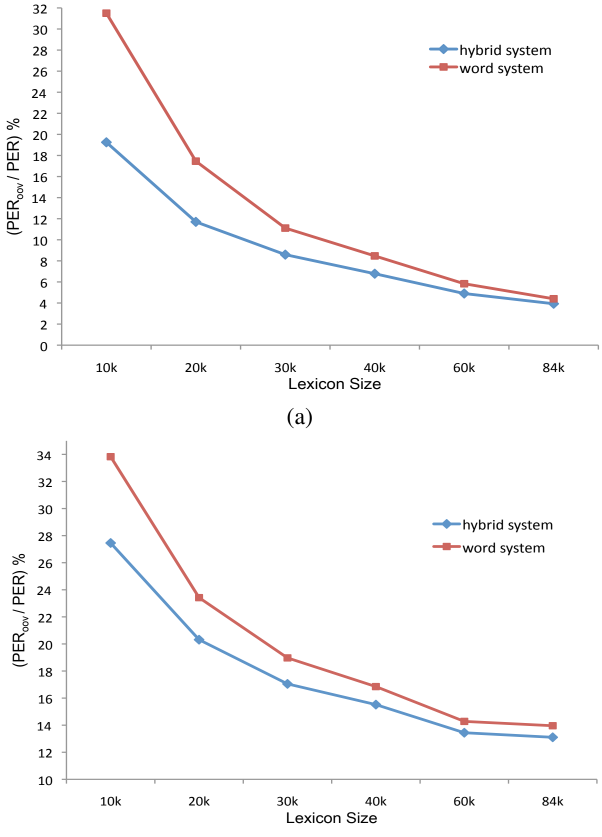
(b)

Figure 1: PER Results: (a) RT-04 (b) MIT Lectures

In order to confirm that the improvements with the hybrid system are indeed in the OOV regions which are now modeled better by the fragments, we computed the phone error rate due to OOV regions (PER_{oov}) for both test sets. The OOV regions for each of the lexicon sizes were derived from the aligned reference transcripts. As one would expect, Figure 2 illustrates that the fraction of errors in the OOV regions for the hybrid systems is consistently lower than word-only systems. This reiterates the idea that a good chunk of the errors due to OOVs can be recovered once these regions can be automatically detected using a hybrid system as an OOV detector (Section 5.2) or directly using the method proposed in Section 6.

5.2. OOV Detection

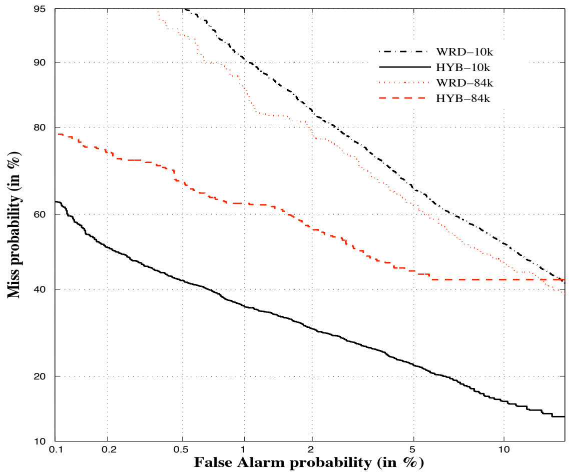
Figure 3 demonstrates the OOV detection results for the word and hybrid systems for lexicon sizes ($10K$ and $84K$). As described in Section 3, OOVs are detected using the posterior probability of fragments in the confusion networks produced by the hybrid system. Word entropy is used to detect OOVs from the confusion networks produced by the word system [15]. The DET curves for the intermediate lexicon sizes for both the hybrid and word systems show the same trend. It is clear that the the hybrid system is consistently better in detecting OOV regions. This coupled with the fact that the hybrid system has a better PER, allows for an increased recovery of OOV term(s), as presented in Section 6.



(a)

(b)

Figure 2: PER in OOV regions as a percentage of the overall PER: (a) RT-04 (b) MIT Lectures



(a)

Figure 3: OOV detection results on RT-04 test set using word and hybrid systems for $10K$ and $84K$ lexicons (words)

6. From Sub-word units to Words

For our goal of recovering OOV words from the output lattices of a word/hybrid automatic speech recognition (ASR) system we need to rely on the phonetic information in the lattice. One method of recovering OOV words uses FSM operations on the lattice to first convert it to a phonetic representation and then recover the OOV words using a enhanced vocabulary dictionary and language model. In this scheme, the first step towards recovering OOV words, is to convert, word and hybrid lattices (L) to phone lattices using the pronunciations in the ASR's lexicon (d). The phone lattices are then composed with an inverted dictionary (D_{inv}) to produce word lattices that are then composed with a large word based language model(W). Both the inverted dictionary D_{inv} and the word model W have a larger vocabulary which may include some of the OOV words of interest. In terms of FSM operations we can express these operations as

$$L \otimes d \otimes D_{inv} \otimes W$$

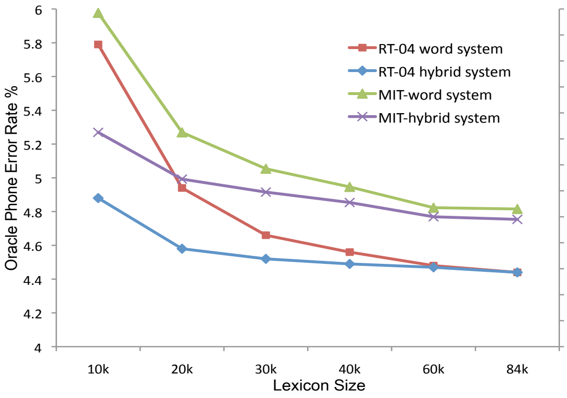
This large dictionary can be obtained in a variety of ways including the world-wide web [4]. In this paper, we used the largest dictionary of size $84K$ to illustrate the effect on WER. Similarly, the large LM used to re-score the word lattices, can be built from several corpora, and we restricted our experiments to the best word LM described in Section 4. The results of this back-transduction are presented in Table 3 for the different hybrid and word systems described in Section 3.

Vocab. Size	10k	20k	30k	40k	60k	84k
Hybrid (%)	15.5	14.9	14.6	14.4	14.2	14.1
Word(%)	17.1	16	15.1	14.6	14.3	14.1

Table 3: WER on the RT-04 Eval set after back-transduction

The hybrid systems are able to produce better PER lattices that result in a reduction in WER. With a $10K$ lexicon, the WER of the hybrid system is 15.5% while that of the word system is at 17.1% and both systems converge to a WER 14.1% when the OOV rate is 0.5%. This back-transduction step opens new avenues for recovering errors from OOV regions.

Oracle phone error rates computed from the confusion networks of the word and hybrid systems (Figure 4) show that the hybrid lattices are richer in phonetic content that can be exploited for deriving improved lexicons.



(a)

Figure 4: Oracle PER of word/hybrid systems on RT-04, shown on the left Y-axis and the MIT data set shown on the right Y-axis

7. Conclusions

In this paper we have demonstrated that the hybrid word/subword systems have a higher phone accuracy, OOV detection performance and better recovery of OOV terms. At an OOV rate of 2.5% on RT-04 we observed a 8% relative improvement in phone error rate (PER), 7.3% relative improvement in oracle PER and 7% relative improvement in WER after recovering the OOV terms. For the various vocabulary sizes we experimented with, the phone recognition accuracy of the hybrid system was better than the word system, on in-domain and out-of-domain test sets. The reduction in PER is also accompanied by a reduction in oracle PER. This helps in better recovery of OOV words from the lattices generated by the hybrid system. Currently, we are exploring approaches to restrict the FSM-based recovery to regions detected as OOVs using fragment posteriors. In the future, we plan to expand this research to improve the ranking of the generated OOV candidates using higher level meta-information derived from the acoustic features of the ASR output and information extracted from the web. The results obtained thus far imply that hybrid systems offer a good paradigm for building open vocabulary systems. It is important to note that the hybrid system serves to enrich the performance of a word-based system particularly in domains with limited training data, and can also serve as a good means to bootstrap better LVCSR systems. However, the hybrid system is not meant to replace a state-of-the-art word-based LVCSR systems but can enhance such its performance for the applications considered here.

8. References

[1] Mark Clements and Marsal Gavald, "Voice/audio information retrieval: Minimizing the need for human ears," in *Proc. ASRU*, 2007.

[2] J. Mamou, B. Ramabhadran, and O. Siohan, "Vocabulary independent spoken term detection," in *Proc. of ACM SIGIR*, 2007.

[3] M. Saraclar and R. Sproat, "Lattice-based search for spoken utterance retrieval," in *Proc. HLT-NAACL*, 2004.

[4] D. Can, E. Cooper, A. Sethy, B. Ramabhadran, M. Saraclar, and C. White, "Effect of pronunciations on oov queries in spoken term detection," in *Proc. ICASSP*, 2009.

[5] D. Vergyri, I. Shafran, A. Stolcke, R. R. Gadde, M. Akbacak, B. Roark, and W. Wang, "The SRI/OGI 2006 spoken term detection system," in *Proc. Interspeech*, 2007.

[6] P. Schwarz, P. Matejka and J. Cernocky, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. ICASSP*, 2006.

[7] O. Siohan and M. Bacchiani, "Fast vocabulary-independent audio search using path-based graph indexing," in *Proc. Interspeech*, 2005.

[8] I. Bazzi, "Modeling out-of-vocabulary words for robust speech recognition," Ph.D. dissertation, MIT, 2002.

[9] A. Stolcke, "Entropy-based pruning of backoff language models," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[10] Stanley F. Chen, "Conditional and joint models for grapheme-to-phoneme conversion," in *Proc. Eurospeech*, 2003.

[11] A. Rastrow, A. Sethy and B. Ramabhadran, "A new method for OOV detection using hybrid word/fragment system," in *Proc. ICASSP*, 2009.

[12] L. Mangu, E. Brill and A. Stolcke, "Finding consensus among words: Lattice-based word error minimization," in *Proc. Eurospeech*, 1999.

[13] S. Chen, B. Kingsbury, L. Mangu, D. Povey, G. Saon, H. Soltau, and G. Zweig, "Advances in speech transcription at IBM under the DARPA EARS program," *IEEE Transactions on Audio, Speech and Language Processing*, pp. 1596–1608, 2006.

[14] J. Glass, T. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent progress in MIT spoken lecture processing project," in *Proc. Interspeech*, 2007.

[15] L. Burget, P. Schwarz, P. Matejka, M. Hannemann, A. Rastrow, C. M. White, S. Khudanpur, H. Hermans, and J. Cernock, "Combination of strongly and weakly constrained recognizers for reliable detection of OOVs," in *Proc. ICASSP*, 2008.