OOV Detection and Recovery using Hybrid Models with Different Fragments

Long Qin, Ming Sun, Alexander Rudnicky

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA
{lqin, mings, Alex.Rudnicky}@cs.cmu.edu

Abstract

In this paper, we address the out-of-vocabulary (OOV) detection and recovery problem by developing three different fragment-word hybrid systems. A fragment language model (LM) and a word LM were trained separately and then combined into a single hybrid LM. Using this hybrid model, the recognizer can recognize any OOVs as fragment sequences. Different types of fragments, such as phones, subwords, and graphones were tested and compared on the WSJ 5k and 20k evaluation sets. The experiment results show that the subword and graphone hybrid systems perform better than the phone hybrid system in both 5k and 20k tasks. Furthermore, given less training data, the subword hybrid system is more preferable than the graphone hybrid system.

Index Terms: OOV detection and recovery, hybrid model, phone, subword, graphone

1. Introduction

Most speech recognition systems are closed-vocabulary recognizers and cannot deal with out-of-vocabulary (OOV) words. On average, one OOV introduces 1.2 word errors [1]. Also, OOV words are usually content words, such as names, locations, etc. Therefore, it is important to develop a speech recognition system which can detect and recover OOVs.

There are several approaches for detecting OOVs: 1) use a hybrid language model (LM) during decoding to explicitly represent OOVs with phones, subwords, graphones or generic word model [2][3][4][5]; 2) use confidence scores and other evidences to locate possible OOV regions [6][7][8][9]; and 3) combine hybrid LM with confidence metrics to further improve OOV detection performance [10][11]. With respect to OOV recovery, phoneme-grapheme alignment can be used to restore the written form of an OOV word [4][12]. Other approaches have used finite state transducers (FSTs) or worked in the spoken term detection framework [13][14][15].

In this paper, we report the performance of OOV detection and recovery using the fragment-word hybrid LM, similar to the method described in [2]. Different fragment types have been proposed, we examine three, the phone, subword, and graphone using hybrid systems, on the WSJ 5k and 20k evaluation sets. We report the word-level recall and precision in the OOV detection task and word error rate (WER) in the recovery task.

The remainder of this paper is organized as follows. Section 2 describes the details of the OOV detection and recovery methods. Sections 3 and 4 discuss our experiments and results. Concluding remarks are provided in Section 5.

2. Method

2.1. OOV detection

A fragment-hybrid LM is applied during decoding to detect the presence of OOVs. We trained an open-vocabulary word LM from a large text corpus and a closed-vocabulary fragment LM from the pronunciations of all words in a dictionary. When training the word LM, all OOVs were matched to the same unknown token "$\langle unk\rangle$". Then by combining the word LM with the fragment LM, a single fragment-hybrid LM was generated. For example, the unigram probability of a fragment in the hybrid LM, $P_H(f_i)$, is calculated as

$$P_H(f_i) = P_W(\langle unk\rangle) \cdot P_F(f_i) \cdot C_{OOV}, \tag{1}$$

where $P_W(\langle unk\rangle)$ is the unigram probability of the unknown token in the word LM, $P_F(f_i)$ is the unigram probability of the fragment in the fragment LM, and C_{OOV} is the cost of entering an OOV word during decoding. Similarly, we can compute N-gram probabilities in the hybrid LM.

2.2. Fragments

We investigate three different types of fragments, phones, subwords, and graphones for suitability in OOV detection and recovery. Phone and subword only model the phonetic level; graphone also considers orthography.

2.2.1. *Phone*

In the phone hybrid system, an N-gram phone LM was trained and combined with a word LM to generate the hybrid LM. Then during decoding, OOVs were represented by phone sequences. For example, our system recognized the OOV word "ashland" as "AE SH AH N", which is close to the correct pronunciation.

2.2.2. *Subword*

Subwords, such as "AH_N" and "EY_SH_AH_N", are iteratively trained phone sequences of variable length [3] as follows: First we add all phones to a subword inventory to ensure the full coverage of all possible words. Then, for each iteration, the most frequent subword bigram is merged and added to the subword inventory. Its occurrences in the training data are concatenated into a single token. This transformed training data is used in the next iteration. The training procedure ends when the target number of subwords is reached.

2.2.3. *Graphone*

A graphone is a grapheme-phoneme pair of English letters and phones. For example, one possible representation of the word "speech" is

$$\text{speech} = \begin{pmatrix} \text{s} \\ \text{S} \end{pmatrix} \begin{pmatrix} \text{pee} \\ \text{P IY} \end{pmatrix} \begin{pmatrix} \text{ch} \\ \text{CH} \end{pmatrix}.$$

In our system, a trigram joint-sequence model is trained from a dictionary and used to segment in-vocabulary (IV) words into graphone sequences [16]. Then a graphone LM is trained and merged with a word LM to produce the hybrid LM. Tokens have a minimum and maximum number of letters and phones; we used the same range for both letters and phones. The minimum is 1 and the maximum is controlled by a parameter L. Therefore, graphones from 1 up to L letters and phones were allowed.

2.3. OOV Recovery

After OOV detection, appropriate spellings were generated for OOVs using the recognized fragment sequences. For the graphone hybrid system, we can simply concatenate letters from the detected graphone sequence to restore the OOV written form. For the phone and subword hybrid systems, a phoneme-to-grapheme conversion was applied. To achieve good phoneme-to-grapheme conversion performance, we trained a 6-gram joint sequence model with short length units, as suggested in [17].

3. Experiment Setup

3.1. Dataset

We tested our system on the WSJ Nov. 92 5k and 20k evaluation sets [18] using the Sphinx3 decoder. The WSJ0 text corpus was used for word LM training. In particular, the top 5k and 20k words in this text corpus were used as vocabulary, yielding an OOV rate of 2% for both tasks. Then an open-vocabulary 5k-word LM and 20k-word LM were trained. The recognition dictionary was generated using CMUdict (v.0.7a). The fragment LM was trained from the dictionary (i.e., without weighting each word by corpus frequency). Following this step, we built a bigram hybrid LM.

We observed a significant improvement by changing the acoustic model from WSJ-SI84 to WSJ-SI284, accordingly we used the WSJ-SI284 model. WER using the word bigram LM was 9.23% and 12.21% on the 5k and 20k task. This is comparable to the results using the standard Lincoln Lab bigram LM.

3.2. Evaluation method

We use recall and precision defined below to measure the OOV detection performance.

$$Recall = \frac{\#\text{correctly detected OOVs}}{\#\text{OOVs in reference}} \times 100\% \qquad (2)$$

$$Precision = \frac{\#\text{correctly detected OOVs}}{\#\text{OOVs reported}} \times 100\% \qquad (3)$$

We calculated recall and precision at the word level which measures both the presence and positions of OOV words in an utterance since for practical purposes (e.g., in a dialog system), knowing where OOVs are located in an utterance is more valuable than simply knowing that OOVs exist. For OOV recovery, we compared the WER before and after restoring the written forms of OOV words.

4. Results

4.1. Phone hybrid system

The OOV detection performance for the phone hybrid system is shown in Fig. 1. We swept the OOV cost C_{OOV} when generating the hybrid LM to generate the recall-precision curve. We

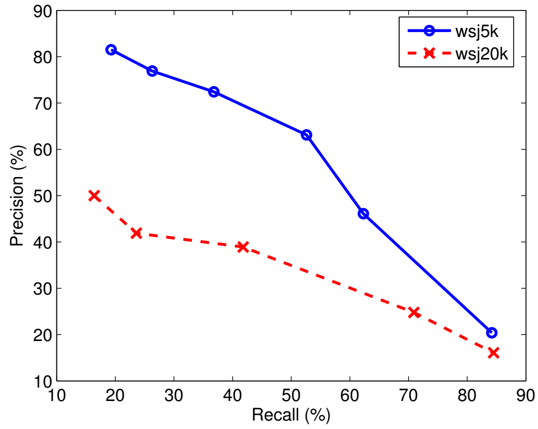

Figure 1: *The OOV detection for the phone hybrid system.*

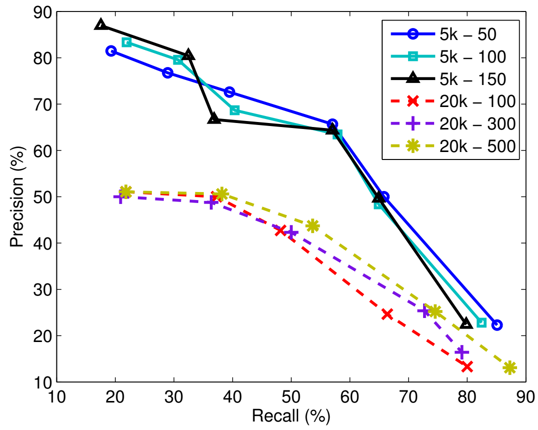

Figure 2: *The OOV detection for the subword hybrid system.*

find that the OOV detection performance for the 5k task is better than for the 20k task. This is due to the much larger dictionary in the 20k system, meaning that an OOV word is more likely to be recognized as IV word. This finding applies to the subword and graphone hybrid systems as well. In the OOV recovery experiment, we observed that the higher the recall, the larger the improvement. But the overall relative improvement after recovery was not as good as expected, because of the high consonant deletion rate in OOV hypotheses (cf. example in Section 2.2.1). Details of OOV recovery results are given in Section 4.4.

4.2. Subword hybrid system

For the subword hybrid system, we investigated OOV detection and recovery performance by varying the number of subwords. For the 5k task, we used 50, 100, and 150 subwords when training the subword LM. Here, 50 subwords means that there are 50 compound subwords plus initial phone subwords. For the 20k task, as there are more IV words for training, 100, 300, and 500 subwords were tested. We find (Fig. 2) that using 50 subwords is better than using 100 or 150 subwords in the 5k task. However, in the 20k task, we achieved the best result with 500 subwords, which is 10 times greater than the optimal number of subwords used in the 5k task. The reason lies in the growth of available training data. There are only 5000 words for training the subword LM in the 5k task, which is not enough to estimate a reliable distribution over subwords. In contrast, in the 20k task, as we had four times more IV words for training, more subwords were affordable. Given a larger dictionary for training, 100 or 150 subwords could be a better choice in the 5k task. In both tasks, more than 150 subwords or 500 subwords were tested, but the performance was worse than the results we presented in Fig. 2.

For OOV recovery, in both the 5k and 20k tasks, we found that the relative improvement of WER after recovery was higher with more subwords. This is reasonable as more subwords requires more iteration runs, thus longer subwords will appear in the hybrid LM. For example, the average length of 50 subwords and 150 subwords in the 5k task is 1.6 and 2.1 phones, while the average length of 100 subwords and 500 subwords in the 20k task is 1.9 and 2.5. Since more consonants will remain in longer subwords during decoding, the phoneme-to-grapheme conversion is more precise.

4.3. Graphone hybrid system

Different graphone lengths were tested to find the optimal graphone set for OOV detection and recovery. In both 5k and 20k tasks, graphones with length of 2, 3 and 4 were applied. Table 1 shows the total number of graphones when varying graphone length. It can be seen that there will be many more graphones if we allow a graphone to have longer sequence of letters and phones. Fig. 3 shows the OOV detection results of graphone hybrid system with different graphone lengths. We can find that the optimal graphone length is 2 in the 5k task and 3 in the 20k task. Again, similar to the subword results, because there is a larger IV dictionary for training in the 20k task, we could use more graphones. For OOV recovery, we observed that better result was gained by using longer graphones. This is consistent with the subword results.

Table 1: *Graphone count over different graphone lengths.*

Graphone Length	2	3	4
WSJ-5K	1167	3000	4122
WSJ-20K	1746	6220	11030

4.4. Comparing the three systems

Fig. 4 presents the OOV detection results of each individual system. In the 5k task, the 50-subword system and the length 2 graphone system were selected. In the 20k task, we chose the 500-subword system and the length 3 graphone system. It can be seen that in the 5k task, the subword hybrid system is better than the phone hybrid system. The reason is that it can utilize a longer history of phones. For instance, a subword bigram "AE_N T" incorporates a history of two phones but any phone bigram only relies on the previous phone. The subword hybrid system is also better than the graphone hybrid system in the 5k task; even if graphone length is as short as 2, there are still more than 1000 graphones. However, given the size of the training corpus, a distribution over such a large inventory cannot be well estimated. So it is not surprising that the graphone hybrid system catches up with the subword hybrid system in the 20k task. Meanwhile, the growth of the training data also leads to a significant improvement over the phone hybrid system relative to the subword and graphone systems.

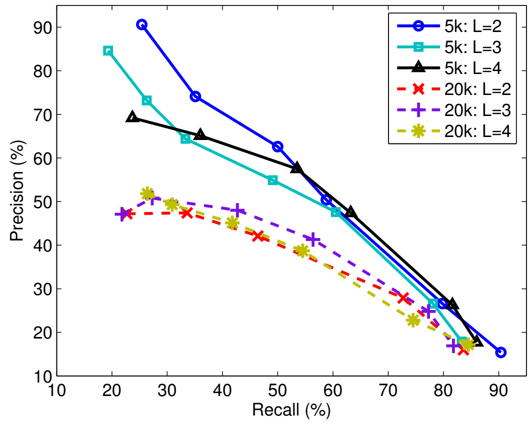

Figure 3: *OOV detection for the graphone hybrid system.*

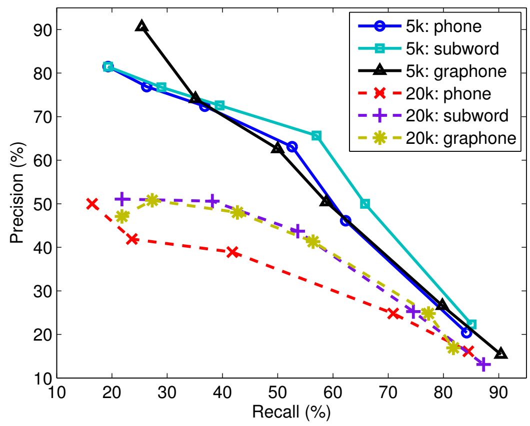

Figure 4: *OOV detection for different systems.*

The OOV recovery results of different systems are given in Fig. 5 and Fig. 6, in which the solid line is the WER before recovery and the dashed line is the WER after recovery. First, we only compare the WER of the three systems before recovery. The subword hybrid system is the best in the 5k task and the graphone hybrid system in the 20k task. And they are both better than the phone hybrid system in either task. We can also learn that when recall is low, the difference among those systems is small. In fact, given low recall, the WER of all the three systems almost remains the same as the baseline word recognition result, sometimes even lower. This implies that we can detect OOVs without affecting WER. Similar to the detection task, with more training data, the advantage of the subword and graphone hybrid systems over the phone hybrid system is more pronounced.

Now, we focus on the OOV recovery performance in Fig. 5 and Fig. 6. We can notice that, as mentioned in each individual system, OOV recovery is more effective when recall is high. Moreover, the best overall performance is obtained by the subword hybrid system in the 5k task and by the graphone hybrid system in the 20k task. In both tasks, we achieved the largest relative improvement after recovery from the graphone hybrid system. This is because in the phone and subword hy-

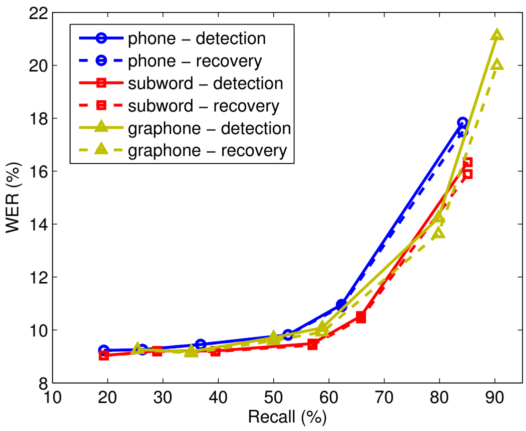

Figure 5: *The OOV recovery results on the 5k task.*

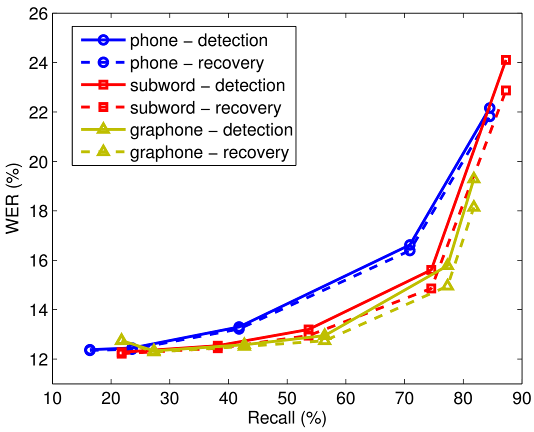

Figure 6: *The OOV recovery results on the 20k task.*

brid systems, the spellings are recovered through the phoneme-to-grapheme conversion, which may involve additional errors. However, in the graphone hybrid system, letters and phones are aligned and simultaneously modeled as a pair. So during decoding, the pronunciation and the written form of an OOV word are found at the same time. Hence, the phoneme-to-grapheme conversion is not necessary.

5. Conclusion

We compared OOV detection and recovery performance for three fragment-word hybrid systems, phone, subword, and graphone. For each system, the configuration, such as the number of subwords or the length of graphone, was varied. We compared performance on two different datasets - the WSJ 5k and 20k tasks. We found that the subword and graphone hybrid systems are significantly better than the phone hybrid system for both OOV detection and recovery tasks. Furthermore, for the subword and graphone hybrid systems, 1) more training data allows more fragments thus better coverage; 2) provided sufficient training data are available, longer subwords or graphones are preferable to shorter ones; 3) the graphone hybrid system provides a larger relative improvement in the recovery task.

6. Acknowledgment

This work was supported in part by the NSF (grants IIS-101273 and IIS-0713441). We thank our reviewers for their comments.

7. References

[1] R. Rosenfeld, "Optimizing lexical and N-gram coverage via judicious use of linguistic data," *Proc. Eurospeech-1995*, pp. 1763-1766, 1995.

[2] I. Bazzi and J. Glass, "Modelling out-of-vocabulary words for robust speech recognition," *Proc. ICSLP-2000*, vol. 1, pp. 401-404, 2000.

[3] D. Klakow, G. Rose, and X. Aubert, "OOV-detection in large vocabulary system using automatically defined word-fragments as fillers," *Proc. Eurospeech-1999*, pp. 49-52, 1999.

[4] M. Bisani and H. Ney, "Open vocabulary speech recognition with flat hybrid models," *Proc. Interspeech-2005*, pp. 725-728, 2005.

[5] T. Schaaf, "Detection of OOV words using generalized word models and a semantic class language model," *Proc. Eurospeech-2001*, pp. 2581-2584, 2001.

[6] F. Wessel, R. Schluter, K.Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 288-298, 2001.

[7] H. Sun, G. Zhang, f. Zheng, and M. Xu, "Using word confidence measure for OOV words detection in a spontaneous spoken dialog system," *Proc. Eurospeech-2003*, pp. 2713-2716, 2003.

[8] H. Lin, J. Bilmes, D. Vergyri, and K. Kirchhoff, "OOV detection by joint word/phone lattice alignment," *Proc. ASRU-2007*, pp. 478-483, 2007.

[9] L. Burget, P. Schwarz, et. al, "Combination of strongly and weakly constrained recognizers for reliable detection of OOVs," *Proc. ICASSP-2008*, pp. 4081-4084, 2008.

[10] A. Rastrow, A. Sethy, and B. Ramabhadran, "A new method for OOV detection using hybrid word/fragment system," *Proc. ICASSP-2009*, pp. 3953-3956, 2009.

[11] C. Parada, M. Dredze, D. Filimonov, and F. Jelinek, "Contextual information improves OOV detection in speech," *Proc. HLT-NAACL-2010*, pp. 216-224, 2010.

[12] K. Vertanen, "Combining open vocabulary recognition and word confusion networks," *Proc. ICASSP-2008*, pp. 4325-4328, 2008.

[13] M. Hannemann, S. Kombrink, M. Karafiat, and L. Burget, "Similarity scoring for recognizing repeated out-of-vocabulary words," *Proc. Interspeech-2010*, pp. 897-900, 2010.

[14] A. Rastrow, A. Sethy, B. Ramabhadran, and F. Jelinek, "Towards using hybrid, word, and fragment units for vocabulary independent LVCSR systems," *Proc. Interspeech-2009*, pp. 1931-1934, 2009.

[15] C. Parada, A. Sethy, M. Dredze, and F. Jelinek, "A spoken term detection framework for recovering out-of-vocabulary words using the web," *Proc. Interspeech-2010*, pp. 1269-1272, 2010.

[16] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, pp. 434-451, 2008.

[17] S.F. Chen, "Conditional and joint models for grapheme-to-phoneme conversion," *Proc. Eurospeech-2003*, pp. 2033-2036, 2003.

[18] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," *Proc. ICSLP-1992*, pp. 899-902, 1992.