OOV Word Detection using Hybrid Models with Mixed Types of Fragments

Long Qin, Alexander Rudnicky

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA

{lqin, air}@cs.cmu.edu

Abstract

This paper presents initial studies to improve the out-of-vocabulary (OOV) word detection performance by using mixed types of fragment units in one hybrid system. Three types of fragment units, subwords, syllables, and graphones, were combined in two different ways to build the hybrid lexicon and language model. The experimental results show that hybrid systems with mixed types of fragment units perform better than hybrid systems using only one type of fragment unit. After comparing the OOV word detection performance with the number and length of fragment units of each system, we proposed future work to better utilize mixed types of fragment units in a hybrid system.

Index Terms: OOV word detection, hybrid model, subword, syllable, graphone, mixed types of fragment units

1. Introduction

Most speech recognition systems are closed-vocabulary recognizers and do not accommodate out-of-vocabulary (OOV) words. But in many applications, e.g., *voice search* or *spoken dialog systems*, OOV words are usually content words such as names and locations which contain crucial information to the success of these tasks. Speech recognition systems in which OOV words can be detected are therefore of great interest.

The fragment-word hybrid speech recognition system applies a hybrid language model (LM) during decoding to explicitly represent OOV words with phones, subwords, graphones, or generic word models [1-4]. Since different hybrid models had been individually proposed, in our previous work we compared the OOV word detection and recovery performance of the phone, subword and graphone hybrid systems [5]. We also studied to utilize complementary hybrid systems by combining multiple systems' outputs using ROVER [6].

In this paper, we investigated system combination for OOV word detection from a different angle by applying the subword, syllable and graphone units simultaneously in one hybrid system. Subwords are phone sequences of variable lengths; syllables are basic phonological "building blocks" of words; while graphones are grapheme-phoneme pairs of letters and phones. Each type of fragment unit has its own advantage and problem. For example, subwords are simple robust units, yet lack linguistic considerations. Syllables maintain phonotactic restrictions to form words, but occasionally produce problematic long rare units. Graphones model both the written form and pronunciation of a word, however, the number of graphone units explodes when allowing longer letter and phone sequences. Therefore, we try to utilize mixed types of units in one system, so that different types of units can complement each other. For this work, we implemented two ways to use mixed types of fragment units in our hybrid system. In the first method, OOV words were divided into three groups and separately modeled by the subword, syllable and graphone units. This is similar to [7], in which researchers used the whole word units together with syllable or morpheme units to model in-vocabulary (IV) words and graphone units for OOV words in an open vocabulary German speech recognition system. In the second method, each occurrence of OOV words was modeled by one type of unit stochastically selected from three types of fragment units. Compared with the first method, where each OOV word is only modeled by one type of fragment unit, in the second method, an OOV word can be modeled by multiple types of fragments if it occurs more than once in the training data. The proposed hybrid systems with mixed types of units were tested on the Wall Street Journal (WSJ) and Broadcast News (BN) datasets.

The remainder of this paper is organized as follows. Section 2 describes the details of hybrid systems using one type of unit and mixed types of units. Sections 3 and 4 discuss experiments and results. Concluding remarks are provided in Section 5.

2. Method

2.1. OOV word detection using a hybrid system

In our hybrid system, we applied a fragment-word hybrid lexicon and LM during decoding to detect the presence of OOV words. The hybrid lexicon was obtained by incorporating fragment units and their pronunciations into the word lexicon. And we trained a flat hybrid LM [8] rather than a hierarchical hybrid LM as in [5]. Specifically, first, the pronunciations of all OOV words were estimated through the grapheme-to-phoneme (G2P) conversion, and then used to train different fragment units, such as subwords, syllables or graphones. After that, OOV words in the training text were replaced by corresponding fragment units to get a new hybrid text corpus. Finally, a hybrid LM was trained from this hybrid text data. When training the hybrid LM, sometimes two OOV words may appear consecutively in the training data. After replacing OOV words with fragment units, the word boundary between two OOV words was lost. To solve this problem, we added two more symbols into the fragment sequence of each OOV word, which are the word start "∧" and word end "$". To achieve good G2P conversion performance, we trained a 6-gram joint sequence model with short length graphone units using the word lexicon [9][10]. We also assigned an OOV cost C_{OOV} to control how likely the system would encounter OOV words during decoding. By tuning C_{OOV}, we can find an optimal configuration of our system to achieve the target OOV word detection performance.

2.2. Fragment units

We built three hybrid systems each with one type of unit, the subword, syllable or graphone units. The subword and syllable systems only model the phonetic level of a word. The graphone system also incorporates the orthography level.

2.2.1. Subword

Subwords, such as "AH_N" and "EY_SH_AH_N$", are iteratively trained phone sequences of variable lengths [2]. First, we initialized the subword inventory with all phones to ensure the full coverage of all possible OOV words. In each iteration, the most frequent subword bigram was merged and added to the subword inventory. Its occurrences in the training data were also concatenated into one single entry. This transformed training data was then used in the next iteration. The training ended when a target number of subword units was reached. In this paper, we used the estimated OOV pronunciations to train subword units.

2.2.2. Syllable

Syllables are often considered as the phonological "building blocks" of words, which can influence the rhythm, prosody and stress of a word. The general structure of a syllable consists of three segments: the onset, nucleus, and coda. The nucleus is normally a vowel or a diphthong, while the onset and coda are usually optional consonants. For example, the word "water" can be split into two syllables: "∧W_AO" and "T_ER$". In our system, we segmented OOV pronunciations into syllables using the Festival lexicon tools [11].

2.2.3. Graphone

A graphone is a grapheme-phoneme pair of English letters and phones. For example, the word "speech" can be represented as

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

To find graphone units, a trigram joint-sequence model was trained from the estimated OOV pronunciations and then used to segment OOV words into grapheme-phoneme pairs [9]. A graphone can have a minimum and maximum number of letters and phones. Here, we used the same range for both letters and phones, where the minimum was set to 1 and the maximum was varied from 2 to 5.

2.3. Hybrid systems with mixed types of units

Since each type of fragment unit has its own advantage, we try to utilize all three types of fragment units in one hybrid system. Here, we compared two methods to use the mixed types of units.

In the first method, hereafter referred as "`mix-hier`", different types of fragment units were combined in a hierarchical way, where we divided OOV words into three groups and used one type of fragment unit for each group. Because the syllable system performed better than the graphone and subword systems, we modeled the most frequent OOV words using syllable units and the less frequent OOV words using graphone and subword units. In particular, we first ranked all OOV words based on their frequencies in the training text. Then, we used syllable units to model the top x percent of OOV words, graphone units to model the following y percent of OOV words, and subword units for the remaining OOV words. The value of x and y were tuned over the development data to get the best OOV word detection performance.

In the second method, hereafter referred as "`mix-flat`", different types of fragment units were combined in a flat way, where each occurrence of OOV words was modeled by one type of fragment unit stochastically selected from subwords, syllables and graphones. We trained each type of fragment unit using the estimated OOV pronunciations. Then, for each OOV occurrence in the training text, we represented it with the fragment units from the stochastically selected unit type. As a result, different from the mix-hier method, here, one OOV word can be modeled by multiple types of fragment units, if it occurs more than once in the training data. The mix-flat system therefore has a better coverage of OOV words from all three types of fragment units. Furthermore, this mix-flat method doesn't require any development data for tuning.

3. Experiment setup

3.1. Dataset

We tested our hybrid systems on the Wall Street Journal (WSJ) and Broadcast News (BN) tasks. The WSJ0 and BN 92-96 text corpora were used to train the bigram hybrid LMs. In both tasks, the 20k most frequent words were chosen as in-vocabulary (IV) words. The acoustic models were trained from the WSJ-SI284 and HUB4-96 BN data. The SPHINX3 decoder was used for recognition [12]. For WSJ, the development data is the sum of the WSJ 92 and 93 20k-word Dev sets, while the evaluation data includes the WSJ 92 20k-word and 93 64k-word Eval sets. For BN, the development and evaluation data are the F0 and F1 sets of the 1996 HUB4 Dev and Eval data. We found that some OOV words in the development and evaluation data were morphological changes of IV words, such as "TANK'S" and "LUMBERING". As such OOV words are not the "new" words our system expects to detect, we manually examined OOV words in the testing data and added those "fake" OOV words into the lexicon. We also removed a few less frequent words from the lexicon to maintain the 20k-word vocabulary size. The final OOV rate of the development and evaluation data for both the WSJ and BN tasks is slightly higher than 2%.

3.2. Evaluation metrics

We used the *miss rate* and *false alarm (FA) rate* defined below to evaluate the OOV word detection performance.

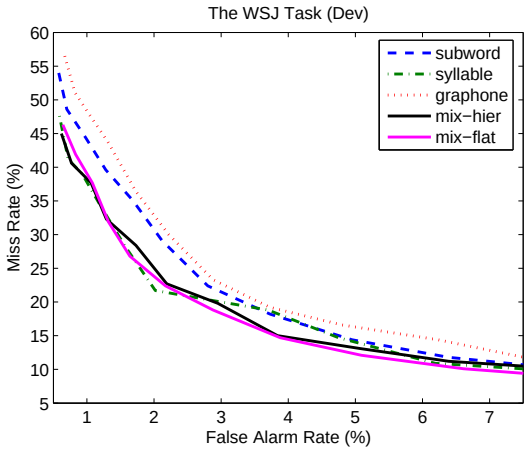
$$Miss = \frac{\#\text{OOVs in reference} - \#\text{OOVs detected}}{\#\text{OOVs in reference}} \times 100\% \quad (1)$$

$$FA = \frac{\#\text{OOVs reported} - \#\text{OOVs detected}}{\#\text{IVs in reference}} \times 100\% \quad (2)$$

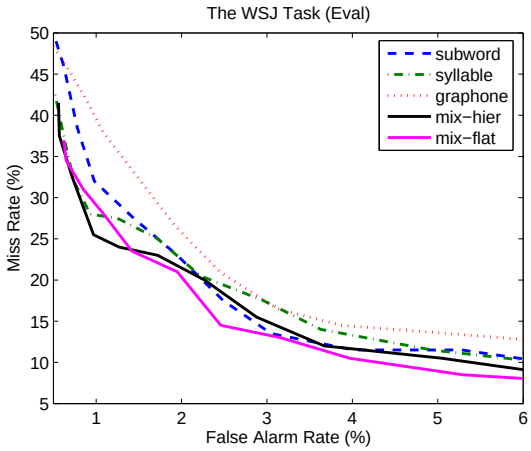
We calculated the *miss rate* and *false alarm rate* at the word level, which measures both the presence and positions of OOV words in an utterance. Because in practical applications, knowing where OOV words are located is more valuable than simply knowing the fact that OOV word(s) exist in an utterance.

4. Experiment results

In our experiments, we tuned x and y for the mix-hier method over the development data. We changed x from 30 to 50 and y from 10 to 50 with a step size of 10. The best performance was achieved when using $x = 40$ and $y = 20$ in both the WSJ and BN tasks. The optimal number of subword units and graphone length for hybrid systems using one type of unit were also tuned on the development data. For WSJ, the 7000-subword system and the length 4 graphone system performed better than others. For BN, we picked the 3000-subword system and the length 3 graphone system. To draw the FA-Miss curve, during decoding, we adjusted the OOV cost C_{OOV} from 0 to 2.5 with a step size of 0.25.

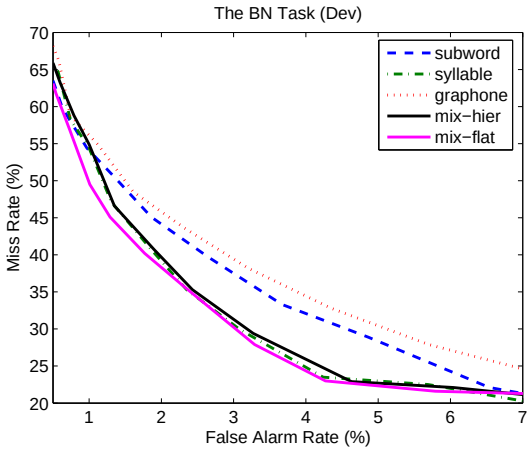


(a) Dev Set

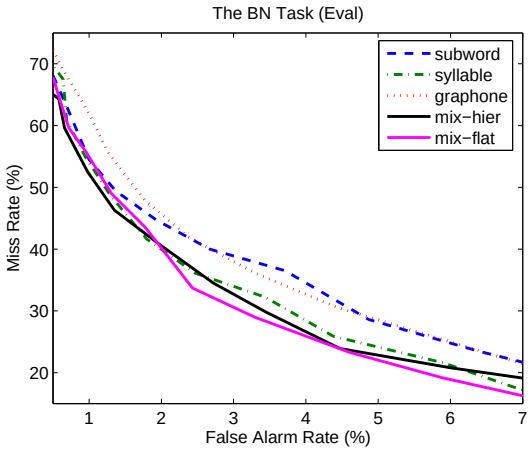


(b) Eval Set

Figure 1: *The OOV word detection results on the WSJ task.*



(a) Dev Set



(b) Eval Set

Figure 2: *The OOV word detection results on the BN task.*

4.1. Hybrid systems with one type of fragment unit

We first compared the OOV word detection performance of hybrid systems with only one type of fragment unit. As shown in Fig. 1 and Fig. 2, we can find that the syllable system and the subword system are better than the graphone system in the WSJ task, while the syllable system is better than the other two systems in the BN task. Although hybrid systems with different fragment units perform differently, do they also complement each other? The distribution of detection errors shared by the subword, syllable and graphone hybrid systems is given in Table 1. If a detection error appears in all three hybrid systems, then it is a common error among "3 systems". On the other hand, if the error only appears in one hybrid system, it is a error in "1 system". We averaged the common errors among different systems from the detection results with $C_{OOV} = [0, 2.5]$ during decoding. It can be seen that across all tests, only about 50% of errors appear in two and three systems. Therefore, if we can perfectly utilize complementary systems, we can then greatly improve the OOV word detection performance.

4.2. Hybrid systems with mixed types of fragment units

The OOV word detection results in Fig. 1 and Fig. 2 show that the mix-flat system usually performs better than the mix-hier system, which may be a benefit of better model of OOV words using all three types of fragment units. Furthermore, by utilizing complementary fragment units, the mix-flat system also outperforms individual hybrid systems with only one type of fragment unit. The mix-hier system can be improved by more carefully tuning the parameters x and y. However, the mix-flat system is still preferable, as it doesn't involve any manual work and development data in training.

Table 1: *The distribution of detection errors shared by the subword, syllable and graphone hybrid systems.*

Errors appear in (%)	WSJ		BN	
	Dev	Eval	Dev	Eval
1 system	51.4	49.6	54.9	53.6
2 systems	25.8	27.0	22.8	24.2
3 systems	22.8	23.4	22.3	22.2

4.3. Results analysis and future work

We used the number of fragment units in a hybrid system to measure the system complexity. If more fragment units are used in a system, it provides a better model of OOV words. But it

also requires more data for training the fragment units and hybrid LM. Table 2 presents the number of fragment units in each hybrid system. To be noticed, as we added the word start and word end symbols, "∧" and "$", into OOV pronunciations, a fragment unit is treated differently depending on where it occurs in an OOV word. As a result, we had a large number of fragment units in our system. Among three systems with one type of unit, the syllable system contains the largest number of units. There are less fragment units in the graphone system. And the subword system is the smallest. For systems using mixed types of units, the mix-hier system has less units than the mix-flat system. In fact, there are more fragment units in the mix-flat system than any other systems. This is because, all three types of fragment units may be used to model the multiple occurrences of one OOV word in the mix-flat system. Hence, the mix-flat system consists of a large portion of units from subwords, syllables as well as graphones.

Table 2: *The number of fragment units in different hybrid systems.*

Task	subword	syllable	graphone	`mix-hier`	`mix-flat`
WSJ	7000	26855	19579	17412	42566
BN	3000	34297	13616	32018	40696

We also calculated the average length of fragment units in each system. The length of a fragment unit is defined as the number of phones it contains. A longer fragment unit contains more phones and is thus easier and more robust for recognition. To account for the probability mess of a fragment unit, the length of a unit is weighted by its frequency in the training data. As given in Table 3, the length of graphone units is the smallest in both tasks. In the WSJ task, all the other systems have a similar unit length. In the BN task, the lengths of syllable and mix-hier units are larger than the lengths of subword and mix-flat units.

Table 3: *The average length of fragment units in different hybrid systems.*

Task	subword	syllable	graphone	`mix-hier`	`mix-flat`
WSJ	2.58	2.54	2.27	2.51	2.50
BN	2.21	2.50	1.94	2.49	2.23

By comparing the OOV word detection performance with the number and length of fragment units, we can get the following observations: (1) The better performance systems are usually built from longer fragment units. Because longer units are more stable, and speech recognition systems recognize longer acoustic units better. This is not true for the mix-flat system in the BN task, where the mix-flat system has a shorter length, but still performs well, which requires our further investigation. (2) The subword system has fewer fragment units than all the other systems, which may indicate that it is a less trained system. Subword units are generated by merging the most frequent phone sequences without any linguistic constraints. It usually produces simpler units than the syllable and graphone systems. On the other hand, syllable units maintain phonotactic restrictions. But the extremely large amount of syllable units possibly causes an overfitting problem in the syllable system; (3) The graphone system performs poorly compared with the other systems. It contains many graphone units, but the unit length is quite small. The length of graphone units can be boosted by allowing longer phone sequnces in a graphone. However, the number of graphone units will also increase dramatically.

Following these observations, we are able to propose a better way to utilize mixed types of fragment units in a hybrid system. We still choose syllable units as the base units. Then for the long rare syllable units, we can divide them into smaller subword units. On the contrary, for the short frequent syllable units, we instead introduce graphone units to distinguish the multiple instances of those units using letters.

5. Conclusion

In this paper, we studied the use of mixed types of fragment units in one hybrid system for OOV word detection. Two hybrid systems with mixed types of fragment units, `mix-hier` and `mix-flat`, were built and compared with systems using one of the following types of units - the subword, syllable and graphone units. From our experimental results, we found that the OOV word detection performance of hybrid systems with mixed types of fragment units are better than systems using only one type of unit. Furthermore, we compared the OOV word detection performance with the number and length of fragment units of each system, and proposed future work to better use different fragment units in a hybrid system.

6. Acknowledgment

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

7. References

[1] I. Bazzi, "Modelling out-of-vocabulary words for robust speech recognition," *Ph.D thesis*, MIT, 2002.

[2] 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.

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

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

[5] L. Qin, M. Sun, and A. Rudnicky, "OOV detection and recovery using hybrid models with different fragments," *Proc. Interspeech-2011*, pp. 1913-1916, 2011.

[6] L. Qin, M. Sun, and A. Rudnicky, "System combination for out-of-vocabulary word detection," to appear in *Proc. ICASSP-2012*, 2012.

[7] M. Shaik, A. El-Desoky, R. Schluter, and H. Ney, "Hybrid language model using mixed types of sub-lexical units for open vocabulary German LVCSR," *Proc. Interspeech-2011*, pp. 1441-1444, 2011.

[8] L. Galescu, "Recognition of out-of-vocabulary words with sub-lexical language models," *Proc. Eurospeech-2003*, pp. 249 252, 2003.

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

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

[11] A. W. Black, P. Taylor and R. Caley, "The Festival speech synthesis system," University of Edinburgh, 1997.

[12] "CMU SPHINX: the open source toolkit for speech recognition," URL: http;//http://cmusphinx.sourceforge.net/.