The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

Investigating the Use of Mixed-Units Based Modeling for Improving Uyghur Speech Recognition

Pengfei Hu, Shen Huang, Zhiqiang Lv

Tencent Research, Beijing, China

alanpfhu@tencent.com, springhuang@tencent.com,zhiqianglv@tencent.com

Abstract

Uyghur is a highly agglutinative language with a large number of words derived from the same root. For such languages the use of subwords in speech recognition becomes a natural choice, which can solve the OOV issues. However, short units in subword modeling will weaken the constraint of linguistic context. Besides, vowel weakening and reduction occur frequently in Uyghur language, which may lead to high deletion errors for short unit sequence recognition. In this paper, we investigate using mixed units in Uyghur speech recognition. Subwords and whole-words are mixed together to build a hybrid lexicon and language models for recognition. We also introduce an interpolated LM to further improve the performance. Experiment results show that the mixed-unit based modeling do outperform word or subword based modeling. About 10% relative reduction in Word Error Rate and 8% reduction in Character Error Rate have been achieved for test datasets compared with baseline system.

Index Terms: uyghur speech recognition, subword, mixed-units,language model

1. Introduction

Uyghur is a Turkic language which is widely used in Western China by Uyghur people. It is an agglutinative language in which words are formed by productive affixation of derivational and inflectional suffixes to a root[1-2]. In most large vocabulary continuous speech recognition systems, the recognition vocabulary consists of a list of word forms observed in the training text, and n-gram language models are estimated over sequences of words. However, in agglutinative languages such as Uyghur, words are relatively long, and the vocabulary size of these languages is growing up proportionally with the corpus size[3]. It is impossible or infeasible to include all the word forms in a lexicon when implementing ASR system. So word based approach will lead to high OOV rates and cause data sparsity issue[4]. Therefore, subword like morphemes is a good choice for these languages, which can be properly combined to produce a wide range of words achieving better lexical coverage. Actually subwords has been conventionally adopted in many inflectional languages, such as Japanese, Korean, Turkish, Finnish, German and Arabic [5-14].

One of the main issues of subword language model is the proper choice of the subword type. For Uyghur speech recognition, the most popular unit is morpheme which is the smallest linguistic component of word that has a semantic meaning. Normally, morphemes are generated from the whole-words by applying word decomposition based on supervised or unsupervised approaches. The supervised approaches make use of linguistic knowledge like in [15], where a set of manual rules are developed. On the other hand, the unsupervised approaches are statistical based data driven approach. The unsupervised approaches are language independent as they do not require any language specific knowledge and can be applied to any language. Another type of subword is the syllable. Syllables in traditional Uyghur language are regular and the general format is CV[CC] (C stands for consonant, V stands for vowel). Although syllables are used as units for various languages like Chinese[16] and Polish[17], there is little related work for the Uyghur language, maybe it is because the Uyghur language includes considerable amount loanwords from Chinese, Arabic, Persian and Russian. In Uyghur speech recognition, the morpheme is mainly investigated as subword unit, since it can provide high coverage, low vocabulary size, and semantic and syntactic relations.

Although subword modeling handles OOV issuses and achieves better performance in Uyghur speech recognition system, it has its own demerits. Compared with whole-word units, the subword units are short, often consisting of one or two phonemes, thus they are more likely to be confused in ASR than word units. Especially, vowel weakening and reduction in Uyghur degrade the effect of subword models. However, compared with subwords, whole-word units provide longer and better linguistic constraint, while causing OOV and data sparseness problems. Therefore, we investigate the use of mixed units in lexicon and language model for improving Uyghur ASR system and explore optimal strategy to make subwords and whole-words bring complementary advantages.

Besides, when generating mixed lexicon finite state transducer(FST) we take careful consideration of some issues related with subword, such as silence modeling and position-dependent phones, which enhances subwords' restrictions and bring a performance gain. Moreover, we introduce an interpolated LM of whole-words and mixed units to reduce CER and WER further. Taken together, mixed units and optimization, along with interpolated LM, allow us to improve open vocabulary Uyghur ASR system from CER of 17.23% to 15.7%. Although mixed units based modeling is applied to speech recognition task in many works[18,19], even in Connectionist Temporal Classification[20], to the best of our knowledge, the related optimization work was not explored for Uyghur speech recognition before.

The remainder of this paper is organized as follows: First we discuss related details of subword based approach in Section 2. Then, the mixed-units based method used in our system is given in Section 3. Next, we demonstrate experimental setup and evaluations in Section 4 and Section 5. Finally, we present the conclusions in Section 6.

2. Subword model

Compared with the word based approach, the subword system's most parts are almost same. For example, whether the units are words or subwords, the n-gram modeling tool creates a model that predicts probability of the next token based on the previous

10.21437/SLTU.2018-45

n-1 ones. But at first we need split whole-words into several subwords, and then substitute the corresponding whole-words in text corpus with the subwords sequence. In this section we will talk about the related details of subword models.

2.1. Subword segmentation methods

The first step for subword system is to define the subword unit. There are many segmentation methods for this, which include supervised and unsupervised ways. Because the supervised methods need linguistic knowledge and lacks flexibility, we investigate two unsupervised data-driven segmentation methods in this work: Morfessor and byte pair encoding(BPE). Intuitively, in both methods the substrings occurring frequently enough in several different word forms are proposed as subwords and the words are then represented as a concatenation of subwords. However, they are based on different principles.

2.1.1. Morfessor

Morfessor is an data-driven method for the segmentation of words into morpheme like units based on the Minimum Description Length(MDL) principle[21]. It is considered a general model for unsupervised induction of morphology from raw text. The general idea behind the Morfessor model is to discover as compact a description of the input text data as possible and It will find those units of language that resemble the surface forms of morphemes. In many works[6,12], The Morfessor has been successfully applied for segmenting Finish, German and other agglutinative languages for speech recognition.

2.1.2. Byte pair encoding

Byte pair encoding is a simple universal text compression scheme, in which the most common pair of consecutive bytes of data is replaced with a byte that does not occur within that data. However, in recent years it is recently popularized for segmenting text in many natural language processing domains such as machine translation[22].The BPE algorithm starts with an initial vocabulary: the characters in the text corpus. The vocabulary is updated using an iterative greedy algorithm. In every iteration, the most frequent bigram (based on current vocabulary) in the corpus is added to the vocabulary by the merging operation. The corpus is again encoded using the updated vocabulary and this process is repeated for a pre-determined number of operations. The number of merging operations is the only hyperparameter to the system which needs to be tuned.

As we know, Uyghur text is written as pronounced and each phoneme is recorded by a character. The vowel assimilation and vowel weakening in Uyghur language brings many informal words in text corpus. In order to avoid irregular words that are harmful to the training process, we train Morfessor and BPE models using a list of words in a prepared vocabulary, which includes most words that occur more than 50 times in the LM training corpus. Nevertheless, for segmentation the trained model will be applied to decompose all of the words, including unseen words.

2.2. Boundary markers

Regardless of the chosen subword units, it is important to be able to reconstruct words from the subwords to produce readable text, which should be done in ASR system's post-processing module. Several boundary markers are explored in [23]. All these marking satisfy the requirement that the word text can be reconstructed in a trivial manner. The actual boundary tag can be changed without any loss of generalization. We will compare left-mark, right mark, left-right-mark boundary markers, which mark the subwords by their location in a word. In left-marked style, a subword is prefix with a character to indicate that that there was no word boundary directly preceding the subword. In right-marked style, a suffix marker is added to a subword if there is no word boundary after it. The left-right-marked style applies markers on both sides of the subwords.

Table 1: *three boudary markers*

mark style	example
left-marked(+m)	$teach\ +er\ +s$
right-marked(m+)	$teach+\ er+\ s$
left-right-marked(+m+)	$teach+\ +er+\ +s$

2.3. Lexicon FST

After the words are splited into subwords, we need to build a pronunciation lexicon for new units. Fortunately in Uyghur language, the spelling of a word indicates its pronunciation. It has almost one-to-one letter-to-phone mapping(except the letter "v" in Latin Uyghur, which has no pronunciation). So we can use the graphemes as phonemes and build the lexicon easily. However, there are three places to be noted when generating lexicon FST for subword using kaldi toolkit[24].

The first is about silence model. The silence is often optionally allowed on word boundaries, but not in the middle of words. Therefore, to avoid to recognize the silence between concatenated subwords, a correct subword implementation needs to be able to indicate what transitions between tokens are actual word boundaries.

The second is about position-dependent phones. In Kaldi, there are four separate phones generated from every original phone, each labeled with its location in the word. This results in labels for the begin, end, internal and single phones. If there is enough data for each of the labeled phones, they are modeled separately, otherwise they will be clustered together during the creation of the decision tree. For position-dependent phones it has to be known if a subword is preceded or succeeded by a word-boundary, information which is not available at the moment the plain-text lexicon is created.

Finally, with all different style subword markings there are restrictions on the possible output sequences the recognizer can generate. The first subword should be a starting subword and the last one an ending subword. For example, the sentence should not start with "+m" marked subwords. To take care of the issues above, we modified the lexicon FST in the same way of [23] to improve the system.

3. Mixed units based system

As shown in many works, a subword based language model can perform better than a whole-word based language model. However, obtaining considerable improvements in our experiment seems hard for naive subword modeling. Enhancing the restrictions of subwords can improve the performance. we will explore the mixed units of whole-words and subwords further in this section, which can make up for subword's deficiencies to some extent.

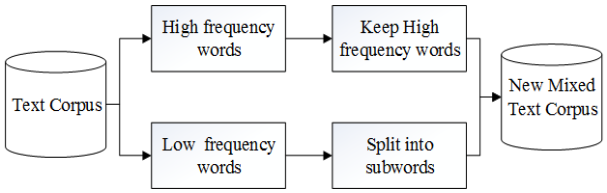

Figure 1: *new corpus on mixed units.*

3.1. Mixed units based Lexicon

For mixed units method, the recognition vocabulary is divided into two parts: The N most frequent words are kept as whole-words without segmentation while the rest of vocabulary consists of sub-words, which come from the segmentaion of low frequency words. It is hoped to prevent the most frequent words from being mixed-up with other sub-words in the search space.

Besides, when generating lexicon FST, we take care of silence model and position-depedent phones for subwords part and will modify the lexicon FST to enhance the restriction of subword locations.

3.2. Mixed units based LM

To train mixed units based LM, we need to transfer the text corpus to mixed units based text. As described in Figure 1, the low frequency words in the text corpus will be substituted by the corresponding subwords sequences and high frequency words are kept unchanged. Language model can be trained based on the new corpus. The mixed units method inherits the subword's merits, and avoid the data sparsity in language modeling.

3.3. LM interpolation

Although subword based language model has better perplexity than whole-words based LM, adding the high frequency words into vocabulary is helpful to recognition performance improvement. So maybe whole-word based LM can offer further assistance. We try to interpolate the mixed unit based LM with whole-words LM and find it can reduce the WER further. When interpolating LM, mixture weights are selected for the LMs that minimize the perplexity of the text corpus and new mixed corpus.

4. Experimental setup

Uyghur is an alphabetic writing language and its standard phoneme set has 32 phonemes, including 8 vowels and 24 consonants. However, when vowels is in the beginning of the words, the letter "v" will be added before the vowels in writing form without any pronunciation. To facilitate language model training, we add 8 v+vowels into phone set such as "va" and "vo". So finally, the phone set in experiment has 40 phonemes. An alternative way to handle letter "v" is considering it as single phone. According to our experiments about two phone sets, they have no big difference in performance. After defining phone set, the lexicon file with 140k words is prepared for model training and decoding.

For acoustic model training, we collected 1000 hours Uyghur speech corpus from several data company. A TDNN-LSTM network with LF-MMI objective is trained based on the collected data using Kaldi toolkit[25]. All experiments use 40-dimensional log-Mel features and 100 dimensional i-vector for speaker adaptation of network. As suggested,we add dropout on the LSTMP layers with a dropout proportion. For decoding, the looped decoding with low frame rate is used[26].

The text corpus for language model training includes two parts: the transcription of speech data and text crawled from website. this corpus is used for vocabulary selection for mixed units and to estimate back-off N-gram LMs using modified Kneser-Ney smoothing by the SRILM toolkit[27]

To evaluate the performance, two test sets are prepared for the Uyghur speech recognition task. One is 4hrs news speech set and another is 1hrs talk speech set. We will compare the results of several methods on these two test sets.

5. Experiment Results

In this section, we will present our experiment results. First, we introduce the baseline experiments. Then we present the results of subword approach. Third, we explore the mixed unites based method and LM interpolation.

5.1. Baseline recognition

In Table 2. we show the results of our baseline recognition experiments using traditional whole-word LMs. We consider the system of 140k whole-words as a reference baseline. while the other baselines are listed for comparison purposes.

Table 2: *OOV rate and recognition results using whole-words models(%)*

Voc size	News set			Talk set		
	OOV	CER	WER	OOV	CER	WER
80k	12.12	19.23	37.45	14.58	23.10	47.36
110k	10.27	18.15	35.92	13.0	21.65	45.43
140k	**9.96**	**17.23**	**34.68**	**11.57**	**20.48**	**44.58**

As shown in Table 2,whether the News set or Talk set, with the increasement of vocabulary size, OOVs decrease and the recoginition performance better. This proves that OOV has a great influence on Uyghur speech recognition. The talk test dataset contains more spontaneous speech. So it seems more diffcult to deal with.

5.2. Results on subword models

In this subsection, we will compare the results of kinds of subword models with different segmentation methods and different boundary marker tags. As mentioned before, whether using Morfessor or BPE, we only consider the words in vocabulary of 140k size when training segmentation models. This can avoid irregular words to impact the ability of segmentation models. The size of subword lexicon is around 40k and no further improvement is found when increasing the lexicon size.

Table 3: *Comparison results of subwords segmentation algorithm and marker styles(%)*

	marker	News set		Talk set	
		CER	WER	CER	WER
Morf	+m	**17.02**	**34.36**	**20.35**	**44.21**
	+m+	17.16	34.54	20.46	44.30
	m+	17.04	34.48	20.48	44.35
BPE	+m	**16.78**	**33.12**	**20.40**	**44.28**
	+m+	16.92	33.80	20.52	44.50
	m+	16.83	33.48	20.48	44.43

Table 3 gives the results of naive subword modeling without lexicon FST modification. We can see that the +m style marking of subwords is most effective for all experiments. The +m+ and m+ marking performs little worse than +m. For segmentation methods, both Morfessor and BPE can bring the improvement. However, for News testset, BPE with left boundary markers obtains the best results while Morfessor with left boundary markers gets the best results for Talk testset.

Besides, for all of subword models with +m makers, we modify the lexicon FST to enforce the subword restrictions and prevent the silence on the subword boundaries. the comparison results of with and without lexicon FST modification is shown in Table 4. It seems that enhancing the subword restriction is very helpful to performance. After modifying lexicon FST, The BPE methods with left marker styles achieves the best performance on both testsets.

Table 4: *Results of subwords modeling with lexicon FST modification (%)*

	L-FST modification	News set		Talk set	
		CER	WER	CER	WER
Morf (+m)	No	17.02	34.36	20.35	44.21
Morf (+m)	Yes	16.57	32.58	19.25	43.40
BPE (+m)	No	16.78	33.12	20.40	44.28
BPE (+m)	Yes	**16.17**	**32.08**	**19.16**	**43.20**

We checked the results of subword based system in details. In fact, the improvement of subword model is mainly from handling OOV. But we also found some cases which is recognized correctly by word based models are failed to be handled by subword model, especially on vowel weakening case. It means that there is still room to improve the performance. So we try to mix whole-words and subwords to reduce the occurrence of these errors. This is our motivation to implement mixed units system.

5.3. Results on mixed units models

Table 5: *results using mixed units based model along with other systems(%)*

	News set		Talk set	
	CER	WER	CER	WER
Baseline	17.23	34.68	20.48	44.58
BPE(+m)	16.78	33.12	20.40	44.28
BPE(+m) L-FST	16.17	32.08	19.16	43.20
Mixed units	**15.81**	**30.95**	**18.84**	**42.74**
Interpolated LM	**15.70**	**30.62**	**18.63**	**42.26**

In Table 4. we give the results of our recognition experiments using LMs based on mixed units. The 50k high fequency words keep unsplited and the rest of 1 millions words are splited into subwords. Finally, the mixed lexcion size is 70k, including 50k whole-words and 20k subwords units. As shown in Table, the mixed units based model outperforms subword models on both testsets.

5.4. Interpolated LM

Motivated by the results of mixed units, we try to incorporate whole-words LM with mixed LM together to enhance the performance. The results in last line of Table 4 shows that Interpolated LM bring a slightly improvement again. This result further proves that combine the whole-words and subwords units can enhance the performance of Uyghur speech recognition system. Finally, by applying mixed units and interpolated LM, along with lexicon FST modification, more than 8% relative reduction in CER in achieved on both testset compared baseline system. For WER, about 10% relative reduction is obtained on average of data amount.

6. Conclusions

Although language model based on subwords improve the performance of Uyghur ASR by handling OOV issues, but also bring some errors because of looser constraints. In this work, we explore several approaches to improve the performance of Uyghur ASR system. Firstly the mixed units are used for language modeling, which keeps high frequency words and only splits low frequency words into subwords. For subwords in mixed units, we modify the lexicon FST to enhance its constraint. The experimental results demonstrate that the mixed units modeling along with enhancing subword constraint can reduce the WER and CER significantly. Besides, the interpolation of whole-words LM and mixed units LM is applied further. Finally, we achieve about 10% relative reduction in WER and 8% reduction in CER on test sets.

7. Acknowledgements

We have greatly benefited from discussions with Bojie Hu and Ambyer.

8. References

[1] N. Tursun, W. Silamu, "Large vocabulary continuous speech recognition in Uyghur: data preparation and experimental results",*Proceedings of ISCSLP*, pp.1-4,2008.

[2] Li, Xin and Cai, Shang,"Large vocabulary Uyghur continuous speech recognition based on stems and suffixes", *Proceedings of ISCSLP*,pp.1-4,2010.

[3] M. Ablimit, G. Neubig, M. Mimura, S. Mori, T. Kawahara, A. Hamdulla, "Uyghur Morpheme-based Language Models and ASR," *Proceedings of ICSP*, 2010.

[4] C. Parada, M. Dredze, A. Sethy, and A. Rastrow, "Learning Sub-Word Units for Open Vocabulary Speech Recognition," *Proceedings of ACL*, 2011.

[5] A. Lee, T. Kawahara, and K. Shikano, "Julius – an open source real-time large vocabulary recognition engine," *Proceedings of Eurospeech*, pp. 1691–1694, 2001.

[6] H. Sak, M. Saraclar, T. Gungor, "Morpholexical and Discriminative Language Models for Turkish Automatic Speech Recognition," IEEE Trans. Audio, Speech & Language Processing, vol. 20, no. 8, pp. 2341-2351, 2012.

[7] A. Berton, P. Fetter, and P. Regal-Brietzmann, "Compound words in large-vocabulary German speech recognition systems," *Prococeeding of International Conference on Spoken Language Processing*, vol. 2, pp. 1165-1168.1996.

[8] A. El-Desoky, C. Gollan, D. Rybach, R. Schluter, and H. Ney, "Investigating the use of morphological decomposition and diacritization for improving Arabic LVCSR," *Proceedings of Interspeech*, pp. 2679-2682, 2009

[9] W. Byrne, J. Hajic, P. Ircing, P. Krbec and J. Psutka, "Morpheme based language models for speech recognition of Czech," in Text,Speech and Dialogue, ser. Lecture Notes in Computer Science,vol.1902, pp. 139-162, 2000

[10] R. Ordelman, A. V. Hassen, and F. D. Jong, "Compound decomposition in Dutch large vocabulary speech recognition," in Proc. European Conf. on Speech Communication and Technology, pp. 225-228. 2003,

[11] M. Creutz, T. Hirsimaki, M. Kurimo, A. Puurula, J. Pylkkonen,V. Siivola, M. Varjokallio, E. Arisoy, M. Saraclar, and A. Stolcke,"Morph-based speech recognition and modeling of outof-vocabulary words across languages," *ACM Transactions on Speech and Language Processing,* vol. 5, no.1, 2007.

[12] A. El-Desoky, M. Shaik, R. Schluter, and H. Ney, "Sub-lexical language models for German LVCSR," *Proceeding of IEEE Workshop on Spoken Language Technology*, pp. 159-164, 2010.

[13] J. Kneissler and D. Klakow, "Speech recognition for huge vocabularies by using optimized sub-word units," *Proc. European Conf. on Speech Communication and Technology*, vol.1, pp. 69-72, 2001.

[14] T. Rotovnik, M. S. Maucec, "Large vocabulary continuous speech recognition of an inflected language using stems and endings," *Speech Communication,* vol. 49, no. 6, pp. 537-452, Jun. 2007

[15] M. Ablimit and T. Kawahara, "Morpheme Segmentation and Concatenation Approaches for Uyghur LVCSR", *International Journal of Hybrid Information Technology*, vol. 8, no. 8, pp. 327–342, 2015.

[16] B. Xu, B. Ma, S. Zhang, F. Qu, and T. Huang, "Speaker independent dictation of Chinese speech with 32K vocabulary," *Proceeding fo ICSLP*, vol. 4, pp. 2320-2323,1996.

[17] M. Piotr, "Syllable based language model for large vocabulary continuous speech recognition of polish," in Text, Speech and Dialogue, ser. Lecture Notes in Computer Science, vol. 5246, pp. 397 401, 2008.

[18] M. A. S. Shaik, A. E. Mousa, R. Schluter, and H. Ney, Hybrid Language Models Using Mixed Types of Sub-lexical Units for Open Vocabulary German LVCSR, in *Proceedings of Interspeech,* pp. 1441-1444, 2011.

[19] C.J.Ni,C.Leung,L.Wang,B.Ma, "Unsupervised data selection and word-morph mixed language model for tamil low-resource keyword search", in Proceeding of ICASSP, 2018

[20] J. Li, G. Ye, A. Das, R. Zhao and Y. Gong, "Advancing Acoustic-to-Word CTC Model", in Proceeding of ICASSP, 2018

[21] M. Creutz and K. Lagus, "Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0," Computer and Information Science Helsinki University of Technology, 2005.

[22] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp.1715-1725.2016.

[23] P.Smit,S.Virpioja,M.Kurimo, "Improved subword modeling for WFST-based speech recognition" *Proceedings of Interspeech*, pp.2551-2555, 2017.

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O.Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit", in *IEEE Workshop on Automatic Speech Recognition and Understanding,* 2011.

[25] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI," in *Proceeding of Interspeech,* pp. 2751-2755, 2016.

[26] Golan Pundak and Tara N Sainath, "Lower frame rate neural network acoustic models," in *Proceeding of Interspeech,* pp.22-26.2016.

[27] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proceeding of International Conference on Spoken Language Processing,* vol. 2, pp. 901-904., Sep. 2002,