

# IMPROVING END-TO-END SPEECH RECOGNITION WITH PRONUNCIATION-ASSISTED SUB-WORD MODELING

Hainan Xu, Shuoyang Ding, Shinji Watanabe

Center for Language and Speech Processing,  
Johns Hopkins University,  
3400 N. Charles St,  
Baltimore MD, U.S.A.  
{hxu31, dings, shinjiw}@jhu.edu

## ABSTRACT

Most end-to-end speech recognition systems model text directly as a sequence of characters or sub-words. Current approaches to sub-word extraction only consider character sequence frequencies, which at times produce inferior sub-word segmentation that might lead to erroneous speech recognition output. We propose *pronunciation-assisted sub-word modeling* (PASM), a sub-word extraction method that leverages the pronunciation information of a word. Experiments show that the proposed method can greatly improve upon the character-based baseline, and also outperform commonly used byte-pair encoding methods.

**Index Terms**— end-to-end models, speech recognition, sub-word modeling

## 1. INTRODUCTION

In recent years, end-to-end models have become popular among the speech community. Compared to hybrid-systems that consist of separate pronunciation, acoustic and language models, all of which need to be independently trained, an end-to-end system is a single neural-network which implicitly models all three. Although modular training of those components is possible [1], an end-to-end model is usually jointly optimized during training. Among the different network typologies for end-to-end systems, the attention-based encoder-decoder mechanism has proven to be very successful in a number of tasks, including *automatic speech recognition* (ASR) [2] [3] [4] and neural machine translation [5][6].

Due to lack of a pronunciation dictionary, most end-to-end systems do not model words directly, but instead model the output text sequence in finer units, usually characters.

THE WORK REPORTED HERE WAS CONDUCTED AT THE 2018 FREDERICK JELINEK MEMORIAL SUMMER WORKSHOP ON SPEECH AND LANGUAGE TECHNOLOGIES, AND SUPPORTED BY JOHNS HOPKINS UNIVERSITY WITH UNRESTRICTED GIFTS FROM AMAZON, FACEBOOK, GOOGLE, MICROSOFT AND MITSUBISHI ELECTRIC RESEARCH LABORATORIES.

This is one of the most attractive benefits of an end-to-end system as it greatly reduce the complexities of overall architecture. However, it only works best for languages where there is a strong link between the spelling and the pronunciation, e.g. Spanish. For languages like English, however, this approach might limit the performance of the system, especially when there is no enough data for the system to learn all the subtleties in the language. On the other hand, linguists have developed very sophisticated pronunciation dictionaries of high quality for most languages, which can potentially improve the performance of end-to-end systems [7].

Sub-word representations have recently seen their success in ASR [8]. Using sub-word features has a number of benefits for ASR, in that it can speed up both training and inference, while helping the system better learn the pronunciation patterns of a language. For example, if a sub-word algorithm segments the word “thank” into “th-an-k”, this will make it easier for the ASR system to learn the association between the spelling “th” and the corresponding sound, which is not a concatenation of “t” and “h”. However, it should also be noted that lots of these methods are designed for text processing tasks such as neural machine translation, and thus are only based on word spellings and do not have access to pronunciation information. It is therefore possible for these algorithms to break a word sequence into units that do not imply well-formed correspondence to phonetic units, making it even more difficult to learn the mapping between phonemes and spellings. For example, if a sub-word model sees a lot of “hys” in the data, it might process the word “physics” into “p-hys-ics”, making the association with the “f” phoneme hard to learn. We argue it is far from ideal to directly apply these methods to ASR and improvements should be made to incorporate pronunciation information when determining sub-word segmentation.

This paper is an effort on this direction by utilizing a pronunciation dictionary and an aligner. We call this method *pronunciation-assisted sub-word modeling* (PASM), which adopts `fast_align` [9] to align a pronunciation lexicon

file and use the result to figure out common correspondence between sub-word units and phonetic units. We then use the statistics collected from this correspondence to guide our segmentation process such that it better caters to the need of ASR. The proposed method would work on a variety of languages with known lexicon, and would also work in other tasks, e.g. speech translation.

This paper is organized as follows. In section 2, we describe prior work; in section 3, we give a detail description of our proposed method, followed by section 4, where we report our experiment results. We will conduct an analysis and discussion of the results in section 5 and then talk about future work in section 6.

## 2. RELATED WORK

The use of a pronunciation dictionary is the standard approach in hybrid speech recognition. [10] use the phone-level alignment to generate a probabilistic lexicon and proposed a word-dependent silence model to improve ASR accuracy; for use in end-to-end ASR models, [7] investigated the value of a lexicon in end-to-end ASR. Sub-word methods have a long history of application in a number of language related tasks. [11] used sub-words units in particular for detecting unseen words. [12] used sub-words units in building text-independent speech recognition systems. [13] improved upon sub-word methods in WFST-based speech recognition.

Apart from the application in ASR, the most recent tide of adopting sub-word representations is largely driven by neural machine translation. [5] proposed to use byte-pair encoding (BPE) [14] to build a sub-word dictionary by greedily keep the most frequent co-occurring character sequences. Concurrently, [15] borrow the practice in voice search [16] to segment words into *wordpiece* which maximizes the language model probability. [6] augments the training data with sub-word segmentation sampled from the segmentation lattice, thus increasing the robustness of the system to segmentation ambiguities.

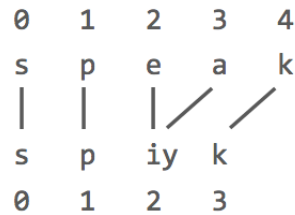
## 3. METHOD

### 3.1. Method Overview

The high-level idea of our method is as follows: instead of generating a sub-word segmentation scheme by collecting spelling statistics from the tokenized text corpus, we collect such statistics only from the *consistent letter-phoneme pairs* extracted from a pronunciation lexicon. The automatically extracted consistent letter-phoneme pairs can be treated as an induced explanation for the pronunciation of each word, and hence, such pairs will ideally contain no letter sequences, i.e. sub-words, that will lead to ill configurations such as “p-hys-ics”.

We generate sub-word segmentation schemes in 3 steps:

**Fig. 1.** A Simple Alignment for the Word “SPEAK”



1. Using an aligner to generate a letter-phoneme alignment from a pronunciation dictionary
2. Extract consistent letter-phoneme pairs from alignment
3. Collect letter-sequence statistics from the consistent letter-phoneme pairs

To simplify the model and generalize to unseen words, we do not perform word-dependent sub-word modeling in this work. Our model generates a list of sub-words with weights, and we split any word with those sub-words.

### 3.2. Method Description

#### 3.2.1. Letter-phoneme Alignment Generation

We use `fast_align` to generate an alignment between letters and phonemes i.e. its pronunciation, which will be able to find common patterns of letter sequences that correspond to certain phonetic units. For example, for the alignment shown in Figure 1, it is represented as a set,

$$\{(0, 0), (1, 1), (2, 2), (3, 2), (4, 3)\}$$

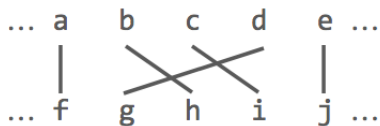
where each element in the set is a pair of (letter index, phone index), both being 0-based. In this case, letters 2 and 3 are aligned to the same phoneme 2. In practice, we could have one-to-one (e.g. “cat”), one-to-many (e.g. “ex”), many-to-one (e.g. “ah”) and even many-to-many alignments (linguistically this should not happen for most languages but this is a good indicator of an “outlier” case, e.g. a French word in an English corpus which the aligner does not know how to process properly).

#### 3.2.2. Finding Consistent Letter-phoneme Pairs

Formally, a consistent letter-phoneme pair  $(L, P)$  is consisted of a letter sequence (or sub-word)  $L = (l_1, \dots, l_n)$  and a phoneme sequence  $P = (p_1, \dots, p_m)$ . These pairs are heuristically extracted from the letter-phoneme alignment generated by `fast_align`, and are then further refined to reduce noise mostly introduced by erroneous alignments.

**Extraction** As `fast_align` is a re-parameterization of IBM model 2, a typical alignment method for statistical machine translation, it does not limit itself in generating

**Fig. 2.** An Alignment with Crossovers



monotonic alignments. There could be cross-overs in its output, like in Figure 2, as well as “null-alignments”, where a letter is aligned to a “null” symbol.

In the case of non-crossing alignments like the one shown in Figure 1, we simply extract each connected sub-sequences. The extracted consistent pairs of this example would be  $(s, s), (p, p), (e-a, iy), (k, k)$ . When there are cross-overs in the generated alignments, like in Figure 2, we take the maximum clustered sub-graph as a consistent pair, i.e. extracting  $(b-c-d, g-h-i)$ .

If a letter is aligned to a “null” symbol, we do not count this as a “cross-over” and keep the letter-to-null mapping for later processing.

**Refinement** Refinement over the consistent letter-phoneme pairs is performed under the following criteria:

1. min-count constraint:  $L$  must occur at least  $N$  times in the training corpus,
2. proportion constraint: of all the words containing  $L$  in the corpus, at least a certain fraction  $p$  of all occurrences is mapped to a particular phone-sequence  $P$ .

### 3.2.3. Collecting Letter Sequence Statistics

Recall that while we use pronunciation lexicon to extract consistent letter-phoneme pairs, our ultimate goal is to collect reliable statistics of the letter sequences (i.e. sub-word) to guide the sub-word segmentation process. Such statistics has nothing to do with phonemes, which means it needs to be marginalized. We perform the marginalization by summing up the counts of each type of letter sequence over all possible types of phoneme sequences. The marginalized counts would act as weights of sub-word units, where higher counts indicate higher weights.

### 3.3. Text Processing

As with all the sub-word modeling methods, our text processing step takes tokenized word sequences as input and segment them into sequences of sub-words. The segmentation process is essentially a search problem operating on the lattice of all possible sub-word segmentation schemes over the word-level input. This segmentation space is constrained by the complete set of sub-words in the segmentation scheme generated above, with hypothesis priorities assigned by the associated weight statistics, where sub-words with higher weights would have

higher priorities. For example, if both “ab” and “bc” are chosen as sub-words, and “ab” occurs more often than “bc” according to the statistics, then “abc” would be split as “ab c” instead of “a bc”.

## 4. EXPERIMENTS

We conduct our experiment using the open-source end-to-end speech recognition toolkit ESPnet [17]. We report the ASR performance on the Wall Street Journal (WSJ) and LibriSpeech (100h) datasets. Our baseline is the standard character-based recipe, using bi-directional LSTMs with projection layers as the encoder, location-based attention, and LSTM decoder, with a CTC-weight of 0.5 during training [4]. To fully see the effect of sub-word methods, we do not perform language model rescoring but report the 1st pass numbers directly.

**Table 1.** WER Results of BPE Systems on WSJ

Num-BPEs	50	108	200	400
dev93	20.7	<b>19.5</b>	21.3	24.6
eval92	15.2	<b>15.6</b>	17.7	20.0

**Table 2.** WER Results on WSJ

	Baseline	PASM	BPE
dev93	20.7	<b>18.5</b>	19.5
eval92	15.2	<b>14.3</b>	15.6

**Table 3.** WER Results on LibriSpeech

	Baseline	BPE	PASM
dev-clean	23.8	29.5	<b>21.4</b>
dev-other	52.8	53.1	<b>50.7</b>
test-clean	23.2	29.5	<b>21.3</b>
test-other	54.8	55.3	<b>52.8</b>

We also compare our systems with BPE baselines. The BPE procedure follows the algorithm described in [5]. All the PASM segmentation schemes are trained using the lexicon included in its default recipe, and we use  $N = 100$  and  $p = 0.5$ . All the other hyper-parameters are independently tuned.

For the WSJ setup, we have kept the number of sub-word units to be the same in BPE and PASM systems (both = 108). The results are shown in Table 2, where we report the word-error-rates on the dev93 and eval92 sets. We see that, the use of BPE improves dev93 performance but hurts performance on eval92. PASM method gives consistent improvements in the 2 datasets.

**Table 4.** Samples of Segmented Text Under the PASM Scheme and BPE Schemes with Various Vocabulary Sizes

Scheme	Text
original	the sale of the hotels is part of holiday's strategy to sell off assets and concentrate on property management
PASM	<u>the</u> <u>sale</u> <u>of</u> <u>the</u> <u>hotels</u> <u>is</u> <u>part</u> <u>of</u> <u>holiday</u> <u>'s</u> <u>str</u> <u>ate</u> <u>gy</u> <u>to</u> <u>sell</u> <u>off</u> <u>as</u> <u>sets</u> <u>and</u> <u>con</u> <u>cent</u> <u>rate</u> <u>on</u> <u>pr</u> <u>o</u> <u>per</u> <u>ty</u> <u>ma</u> <u>na</u> <u>ge</u> <u>me</u> <u>nt</u>
BPE-108	<u>the</u> <u>sale</u> <u>of</u> <u>the</u> <u>hotels</u> <u>is</u> <u>part</u> <u>of</u> <u>holiday</u> <u>'s</u> <u>str</u> <u>ate</u> <u>gy</u> <u>to</u> <u>sell</u> <u>off</u> <u>as</u> <u>sets</u> <u>and</u> <u>con</u> <u>cent</u> <u>rate</u> <u>on</u> <u>pr</u> <u>o</u> <u>per</u> <u>ty</u> <u>ma</u> <u>na</u> <u>ge</u> <u>me</u> <u>nt</u>
BPE-200	<u>the</u> <u>sale</u> <u>of</u> <u>the</u> <u>hotels</u> <u>is</u> <u>part</u> <u>of</u> <u>holiday</u> <u>'s</u> <u>str</u> <u>ate</u> <u>gy</u> <u>to</u> <u>sell</u> <u>off</u> <u>as</u> <u>sets</u> <u>and</u> <u>con</u> <u>cent</u> <u>rate</u> <u>on</u> <u>pr</u> <u>o</u> <u>per</u> <u>ty</u> <u>ma</u> <u>na</u> <u>ge</u> <u>me</u> <u>nt</u>
BPE-400	<u>the</u> <u>sale</u> <u>of</u> <u>the</u> <u>hotels</u> <u>is</u> <u>part</u> <u>of</u> <u>holiday</u> <u>'s</u> <u>str</u> <u>ate</u> <u>gy</u> <u>to</u> <u>sell</u> <u>off</u> <u>as</u> <u>sets</u> <u>and</u> <u>con</u> <u>cent</u> <u>rate</u> <u>on</u> <u>pr</u> <u>o</u> <u>per</u> <u>ty</u> <u>ma</u> <u>na</u> <u>ge</u> <u>me</u> <u>nt</u>

We also report the more BPE results on WSJ, adjusting number of BPE units in Table 1. We can see that having more BPEs actually hurts the performance<sup>1</sup>. This is likely because of the limited data-size of WSJ, which makes it hard to learn reliable BPE units.

In Table 3, we report the WER results on the LibriSpeech dataset, using the parameters described in [8]. We have seen that PASM significantly improves the character-based baseline; BPEs do not help in this case, possibly due to poor hyper-parameter tuning.

## 5. ANALYSIS

In Table 4, we show the output after the BPE procedure of the first sentence in the WSJ training data, and compare that with the result of the PASM algorithm<sup>2</sup>.

From the examples above, we observe the following:

- The PASM method correctly learns linguistic units, including “le”, “th”, “ay”, “ll”, “ll”, “ss”, “ge”, which correspond to only one phoneme, but were not correctly handled in the BPE case.
- The BPE learns some non-linguistic but frequent-seen units in data, e.g. “the”, “ate”. In particular, the pronunciation associated with “ate” in the 2 occurrences are very different (concentr-ate vs str-ate-gy), which might make it harder for the system to learn the associations.
- As the number of BPE units increases, we see more sub-word units that do not conform to linguistic constraints, e.g. “as-s-e-t-s” and “of-f” in BPE-400. In this case, the 2nd “s” “asset” and 2nd “f” in “off”

<sup>1</sup>The character-based baseline has a vocabulary-size of 50.

<sup>2</sup>For clear presentation, we use the underline  character to represent a “start-of-word” symbol.

would have to be silent in terms of pronunciation, which would likely confuse the training of end-to-end systems unless there is a huge amount of data.

## 6. CONCLUSION AND FUTURE WORK

In this work, we propose a sub-word modeling method for end-to-end ASR based on information from their pronunciations. Experiments show that the proposed method gives substantial gains over the letter-based baseline, as measured by word-error-rates. The method also outperforms BPE-based systems. We postulate that the improvement comes from the fact that, the proposed method learns more phonetically meaningful sub-words for speech tasks, unlike BPE which only take the spelling into consideration.

There are a lot of future work directions that we plan to explore. We will design new algorithms for aligning pronunciation dictionaries that is tailored for speech tasks; we will combine the proposed method with BPE to further improve ASR performances and speed up systems; we also plan to investigate the application of the proposed method in hybrid ASR, machine translation, as well as speech translation.

## 7. REFERENCES

- [1] Zhehuai Chen, Qi Liu, Hao Li, and Kai Yu, “On modular training of neural acoustics-to-word model for lvcsr,” *arXiv preprint arXiv:1803.01090*, 2018.
- [2] Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjali Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonnina, et al., “State-of-the-art speech recognition with sequence-to-sequence models,” in *2018 IEEE Interna-*

- tional Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 4774–4778.
- [3] Suyoun Kim, Takaaki Hori, and Shinji Watanabe, “Joint ctc-attention based end-to-end speech recognition using multi-task learning,” in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4835–4839.
- [4] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi, “Hybrid ctc/attention architecture for end-to-end speech recognition,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.
- [5] Rico Sennrich, Barry Haddow, and Alexandra Birch, “Neural machine translation of rare words with subword units,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016.
- [6] Taku Kudo, “Subword regularization: Improving neural network translation models with multiple subword candidates,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, 2018, pp. 66–75.
- [7] Tara N Sainath, Rohit Prabhavalkar, Shankar Kumar, Seungji Lee, Anjali Kannan, David Rybach, Vlad Schogol, Patrick Nguyen, Bo Li, Yonghui Wu, et al., “No need for a lexicon? evaluating the value of the pronunciation lexica in end-to-end models,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5859–5863.
- [8] Albert Zeyer, Kazuki Irie, Ralf Schlüter, and Hermann Ney, “Improved training of end-to-end attention models for speech recognition,” *arXiv preprint arXiv:1805.03294*, 2018.
- [9] Chris Dyer, Victor Chahuneau, and Noah A Smith, “A simple, fast, and effective reparameterization of ibm model 2,” in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 644–648.
- [10] Guoguo Chen, Hainan Xu, Minhua Wu, Daniel Povey, and Sanjeev Khudanpur, “Pronunciation and silence probability modeling for asr,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [11] Ivan Bulyko, José Herrero, Chris Mihelich, and Owen Kimball, “Subword speech recognition for detection of unseen words,” in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.
- [12] Jean-Manuel Van Thong, Pedro Moreno, and Edward Whittaker, “Vocabulary independent speech recognition system and method using subword units,” Feb. 20 2007, US Patent 7,181,398.
- [13] Peter Smit, Sami Virpioja, Mikko Kurimo, et al., “Improved subword modeling for wfst-based speech recognition,” in *INTERSPEECH 2017–18th Annual Conference of the International Speech Communication Association*, 2017.
- [14] Philip Gage, “A new algorithm for data compression,” *The C Users Journal*, vol. 12, no. 2, pp. 23–38, 1994.
- [15] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean, “Google’s neural machine translation system: Bridging the gap between human and machine translation,” *CoRR*, vol. abs/1609.08144, 2016.
- [16] Mike Schuster and Kaisuke Nakajima, “Japanese and korean voice search,” in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2012, Kyoto, Japan, March 25-30, 2012*, 2012, pp. 5149–5152.
- [17] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplín, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al., “Espnet: End-to-end speech processing toolkit,” *arXiv preprint arXiv:1804.00015*, 2018.