

SPELL MY NAME: KEYWORD BOOSTED SPEECH RECOGNITION

Namkyu Jung¹, Geonmin Kim¹, Joon Son Chung²

¹Naver Corporation, South Korea

²Korea Advanced Institute of Science and Technology, South Korea

ABSTRACT

Recognition of uncommon words such as names and technical terminology is important to understanding conversations in context. However, the ability to recognise such words remains a challenge in modern automatic speech recognition (ASR) systems.

In this paper, we propose a simple but powerful ASR decoding method that can better recognise these uncommon keywords, which in turn enables better readability of the results. The method boosts the probabilities of given keywords in a beam search based on acoustic model predictions. The method does not require any training in advance.

We demonstrate the effectiveness of our method on the LibriSpeech test sets and also internal data of real-world conversations. Our method significantly boosts keyword accuracy on the test sets, while maintaining the accuracy of the other words, and as well as providing significant qualitative improvements. This method is applicable to other tasks such as machine translation, or wherever unseen and difficult keywords need to be recognised in beam search.

Index Terms— contextual biasing, speech recognition, keyword boosting, keyword score, beam search.

1. INTRODUCTION

With the advances in deep learning technology, the performance of automatic speech recognition systems have seen tremendous improvements in the recent years [1, 2, 3, 4]. The earlier models have struggled to overcome the overfitting problem given insufficient data, but the development of semi-supervised methods like *wav2vec 2.0* [5] has enabled strong performance with a smaller amount of labeled data. Moreover, with the application of various augmentation methods [6], models have become better at recognising speech in unfamiliar environments or in noisy environments.

Despite this, recognising unseen or uncommon words like person names, location names or technical terminologies is left unsolved. In fact, this problem applies to humans as well. No matter how good a person is at listening, it is almost impossible to understand a conversation full of unknown words. Unfortunately, these words might play a very important role in understanding the conversation, even though the total amount of occurrence might be small. Therefore, we focus on the problem of recognising keywords that were not observed during training.

Contextual information has been researched by using fusion methods with trained language models (LM) [7, 8, 9], but also there are works dealing with *contextual biasing* which utilises a specific context such as named entity or personalised contacts. Majority of works require training an additional representation in advance such as bias encoder [10, 11, 12], bias LM [13], class based LM with finite-state transducer (FST) [14, 15, 16], or additional data augmentation of a named entity with Text-to-speech [17]. Our work is similar to [18] that the bias information is encoded to trie structure. However, they also require additional training

of RNN-T [19] and LSTM-LM [20] to utilise bias information, whereas our method requires only a list of keywords we are interested in.

In short, we assume that there is a list of keywords that might appear in the conversation. Our method promotes them through beam search decoding. We use a character-level Connectionist Temporal Classification Model (CTC) [21] as an ASR model, specifically a pre-trained *wav2vec 2.0* model, but other kinds of model can also be used as long as they are supported by the beam search decoder. We can use the decoding strategy with or without a language model, and in both cases the strategy will prove to be effective. In the real-world, keywords can be a list of characters in a book, a list of people attending a meeting, or a dictionary of technical terminology related to a lecture.

While performing a beam search, the decoding method favours the given keywords if the input speech is pronounced similar to the given keywords. Since the beam search is based on the acoustic model (AM) probabilities, this method would not be activated unless the speech actually contains a keyword or a word with similar pronunciation. The method does not require any training in advance. Our extensive experiments demonstrate that we see a significant boost in keyword accuracy, while maintaining the accuracy on the rest of the output.

2. KEYWORD-BOOSTED DECODING

In this section, we describe the proposed keyword-boosted decoding strategy.

2.1. Keyword Prefix Tree Search

We first prepare a prefix tree (Trie) for the given list of keywords. Prefix tree is to figure out if a prefix is a part of some keyword we are paying attention to. Each node in the tree consists of a token and a keyword index if the path from the root node constitutes a keyword, otherwise a non-keyword index (-1). A keyword can be made of multiple words, but to simplify a problem, we only use single-word keywords.

Figure 1 shows an example of keyword prefix tree with sample keyword set {*cat*, *car*, *coat*}. Each node has its token as a state and a word index. First node represents the root node, n_0 , and the colored nodes are *leaf nodes*. Dashed node means going back to the n_0 since there is no child node.

For each step in the beam search, we will find the next node from the corresponding node and the time complexity for finding next node is $O(1)$. If there are many keywords to be considered, building the trees would be relatively time consuming, but this only needs to be done once, and can be reused thereafter in multiple inferences sharing the same keyword list.

2.2. Keyword-boosted Beam Search

Once a prefix tree of keywords \mathcal{K} is prepared, we can decode the acoustic model output with keyword-boosting algorithm. Let $P_{AM}(s)$ for each

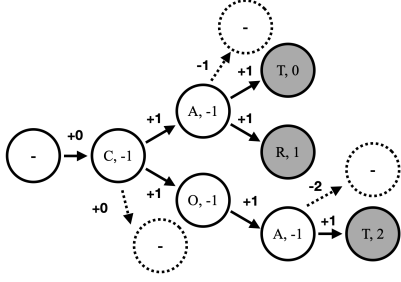


Fig. 1. Keyword prefix tree when a list of keyword consists of ('cat', 'car', 'coat') in character level.

$s \in V$ be a probability of occurring a token s based on an AM where V is the set of tokens. Beam search proceeds along the time step $t \leq T$ with beam width B . For each step t , each b -th beam has a state $s_{t,b}$, which can be a character or a sub-word. Additionally for each beam, node $n_{t,b}$ of the prefix tree \mathcal{K} which begins at the n_0 and *traverse* the tree \mathcal{K} as the beam search proceeds. For CTC models which has *frame-synchronous* decoder [21], $s_{t,b}$ does not always change at every timestep t . Therefore, $n_{t,b}$ would be updated only when its state has actually changed.

After each beam search step and the state $s_{t,b}$ is set, the algorithm updates the current tree node $n_{t,b}$ according to the search result (Step 1). If $s_{t,b}$ is in the children of $n_{t,b}$ we can *traverse* $n_{t,b}$ to the child whose state is $s_{t,b}$. If it is not, $n_{t,b}$ should go back to n_0 . However, there is a chance that $s_{t,b}$ is included in the children of n_0 . In this case, $n_{t,b}$ will traverse to the child of n_0 whose state is $s_{t,b}$.

Before proceeding to the next step $t + 1$, the decoder calculates the keyword boosting score K_t for each candidate next state like a language model score (Step 2). This score intends to give more weights to the children of the current node which continue to go to the next keyword token. We introduce a new hyperparameter *keyword weight* w_k for controlling the strength of the boosting algorithm. The decoder gives w_k for the children of the $n_{t,b}$ and the 0 for the rest of states. Therefore, the keyword score for each b -th beam candidate is

$$K_{t,b}(s) = \begin{cases} w_k, & \text{if } s \in \text{children}_{\mathcal{K}}(n_{t,b}), n_{t,b} \neq n_0 \\ 0, & \text{otherwise,} \end{cases}$$

for each $s \in V$.

Accordingly, even if an AM does not give much probability to a certain state, K_t may give some extra score to boost the state when its node is in the middle of keyword prefix path. Excessively high value of w_k may result in *overboost* (boosting to be a keyword even if it does not actually presented) and a low value may result in nothing but a vanilla beam search. Moreover, this score is not applied from the n_0 because we don't want to boost keywords from the beginning, which may also cause overboost. Detailed algorithm for this method is represented in Algorithm 1.

Using the keyword score defined above, for each time t , the beam search finds B states in the order of maximising the following score,

$$\log P_{AM}(s) + w_{LM} \log P_{LM}(s|H) + w_k K_t(s|H),$$

where H represents the history.

2.3. Keyword subtractive cost

Being on the tree path does not guarantee that the beam is actually containing a keyword. When the path on \mathcal{K} is about to break because

Algorithm 1 Keyword-boosted beam search step at time t .

```

 $K_{t,b}(s) \leftarrow K_{t-1,b}(s), \forall s \in V, \forall b \leq B$ 
for  $b \leq B$  and  $s_{t,b} \neq s_{t-1,b}$  do ▷ update if state has changed.
  * Step 1: Update current node
  if  $s_{t,b} \in \text{children}_{\mathcal{K}}(n_{t-1,b})$  then
     $n_{t,b} \leftarrow \text{traverse}_{\mathcal{K}}(n_{t-1,b}, s_{t,b})$ 
  else if  $s_{t,b} \in \text{children}_{\mathcal{K}}(n_0)$  then
     $n_{t,b} \leftarrow \text{traverse}_{\mathcal{K}}(n_0, s_{t,b})$ 
  else
     $n_{t,b} \leftarrow n_0$ 
  end if
  * Step 2: Calculate keyword score
  if  $n_{t,b} \neq n_{t-1,b}$  then
    if  $n_{t,b} = n_0$  then
       $K_{t,b}(s) \leftarrow 0, \forall s \in V$ 
    else
      if  $n_{t,b} \neq \text{leaf\_node}$  then
         $K_{t,b}(s) \leftarrow -w_k * \text{depth}_{\mathcal{K}}(n_b), \forall s \in V - \{\text{blank}\}$  ▷ subtractive cost
      end if
       $K_{t,b}(s) \leftarrow w_k, \forall s \in \text{children}_{\mathcal{K}}(n_{t,b})$ 
    end if
  end if
end for

```

# keywords	dev-clean	dev-other	test-clean	test-other
1%	714 (1.3)	1,008 (2.0)	750 (1.4)	1,178 (2.3)
5%	2,632 (4.9)	3,842 (7.5)	2,584 (4.9)	4,126 (7.9)

Table 1. The number of keyword occurrence (the ratio to total word occurrence in %). Keywords account for only a very small percentage of total words.

it is not making a keyword, we have to subtract the accumulated value up to the current node. If the next state is not on the tree, we subtract $w_k * \text{depth}_{\mathcal{K}}(n_{t,b})$ as described in Algorithm 1. Note that we do not have to subtract on CTC blank state which is still on the keyword path.

In Figure 1, dashed line represents the escape from the tree which receives the subtractive cost and go back to the n_0 . For example, if the prefix path is *coa* and the next state is going to be *l*, then it will get -2 as a subtractive cost and the node will be initialised to the n_0 .

3. EXPERIMENTS

This section describes the ASR model and the datasets used in the experiments, and the results that demonstrate the effectiveness of our proposed method.

3.1. Baseline ASR system

We use the *wav2vec 2.0 LARGE* [5] model as the baseline ASR model. The model is pre-trained on the unlabeled audio data of LibriVox dataset [22] and fine-tuned on either 100 hours and 960 hours of transcribed LibriSpeech dataset[23]. The two models are selected to represent scenarios with varying amount of labeled in-domain data. We do not fine-tune the data further, but use a 4-gram word LM that is trained on the LibriSpeech LM corpus [24]. In addition, we use a 6-gram character-level LM during decoding, as a multi-level LM similar to [25]. This decoding method essentially use the character-level LM, and then

	$w_{LM}=0.0$				$w_{LM}=1.0$			
	dev		test		dev		test	
	clean	other	clean	other	clean	other	clean	other
100h fine-tuned								
No Boosting	3.15	6.42	3.06	6.10	2.41	4.94	2.44	4.80
1% Keywords	3.08/3.06	6.35/6.33	2.96/ 2.94	6.00/5.95	2.39/ 2.37	4.89/4.88	2.41/ 2.40	4.78/ 4.76
5% Keywords	3.06/ 3.01	6.28/ 6.27	2.95/2.96	5.97/ 5.91	2.39/2.38	4.86 /4.89	2.41/2.42	4.78/4.78
960h fine-tuned								
No Boosting	2.16	4.56	2.13	4.46	1.77	3.51	1.78	3.61
1% Keywords	2.13/2.14	4.49/4.48	2.09/2.10	4.39/4.38	1.74/1.75	3.48/3.48	1.76/1.76	3.59/ 3.57
5% Keywords	2.09 /2.13	4.42 /4.47	2.08 /2.15	4.37 /4.41	1.72 /1.75	3.43 /3.47	1.75 /1.76	3.60/3.61

Table 2. Word Error Rates (WER) on LibriSpeech with and without n-gram LM and keyword boosting $w_k = (0.6/1.2)$.

	LM	✗			✓		
	w_k	P	R	F1	P	R	F1
100h	0.0	98.9	86.0	92.0	99.0	89.5	94.0
fine	0.6	98.9	92.4	95.5	98.9	92.5	95.6
-tuned	1.2	97.5	94.7	96.1	98.7	93.2	95.9
960h	0.0	99.6	94.5	97.0	99.3	95.6	97.4
fine	0.6	99.2	97.7	98.5	99.2	97.9	98.5
-tuned	1.2	97.8	98.5	98.1	98.8	98.4	98.6

Table 3. Precision (P), Recall (R) and F1-score (F1) on the LibriSpeech test-clean with boosting 1% keywords for 100h fine-tuned, 960h fine-tuned model respectively.

substitute character-level probability with word-level ones when each word is made. We will compare the case of LM weight $w_{LM} = 1.0$ or 0.0 (no LM) since the use of LM is optional in our method. Beam search decoding has 100 beams and other fixed model weights.

3.2. Datasets and keyword extraction

LibriSpeech. We evaluate our method on LibriSpeech dev and test sets. On this dataset, we extract the keyword set for each session (book) using the TF-IDF [26] method. LibriSpeech dataset can be classified by the book where it comes from. From the metadata given by the dataset, we can classify every pair of audio/text file from 960h train dataset into the book name and collect data according to its book. The keywords are only extracted from the training set – they are not extracted from the dev and test sets since we have to only use keywords obtainable in advance. With the collected text data for each book, we build a TF-IDF model and extract book-wise keywords with top $n\%$ of the words according to TF-IDF scores. All single-letter word is excluded in the list.

We take two set of keywords from book-wise keywords, top 1% and top 5% of the words according to TF-IDF scores. The 1% set has 16.6 words and the 5% set has 85.5 words for each book on average. We build prefix trees \mathcal{K}^i for each i -th book and put them into every audio file from the corresponding book. Since this tree have the whole list of keywords from the book so the most of them are not occurring in each audio file. Table 1 presents the total number of extracted keywords in each dataset.

In-house dataset. We also perform experiments using Korean-language in-house datasets consisting of *Clova Note* (transcription service) dataset and *NAVER VLive* (Live show for celebrities) dataset. The *Clova Note* dataset has 5,446 audio segments that belong to 365 sessions¹, and each

¹These are purposefully recorded test data, not user-uploaded data.

channel has 7.5 keywords in average extracted by humans. The *NAVER VLive* dataset has 17 audio recordings spoken by two K-pop groups (*BTS*, *Blackpink*), and for each audio, the keyword list contains the members' name (real name and stage name) of each group, 25 words for *BTS* and 11 words for *Blackpink*. The baseline model used in this experiment is also a *wav2vec 2.0*-based model, which has been trained on general-domain Korean language data.

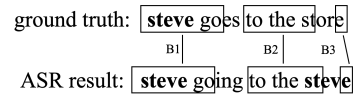


Fig. 2. Matching blocks between the ground truth and ASR result. Bold text represents a keyword.

3.3. Metric

Since the number of keywords is relatively small compared to the size of the corpus (Table 1), the improvement in character error rates (CER) and word error rates (WER) would be limited even if our method is clearly effective in recognising the keywords.

We use `python difflib`² for string comparison which enables us to compare two strings by matching blocks between them. By comparing between the ground truth and the ASR result with the boosting keywords, we can count the *true positive* (TP): the number of keywords in a matching block, *false positive* (FP): the number of keywords in the ASR result but not in a matching block, *false negative* (FN): the number of keywords in the ground truth but not in a matching block.

For example in Figure 2 with 3 matching blocks, we have TP equals to 1 because the keyword **steve** is in the first matching block (B1) and FP also equals to 1 because the second **steve** in ASR result is not inside of any blocks. For a similar reason, FN equals to 0. Consequently, precision is 50%, recall 100% and F1-score 66.7%.

Based on these, we can calculate precision, recall and F1-score with respect to the keywords. Our work focuses on increasing recall while minimizing reduction in precision.

3.4. Results

LibriSpeech. In Table 2, we report the results on LibriSpeech datasets with LM. Upper section is the results with 100h fine-tuned model representing the low-resource environment, and the lower section with 960h fine-tuned model representing full-resource environment. The first row of each section shows the result without keyword boosting, the

²<https://docs.python.org/3/library/difflib.html>

Positive results	keywords	milner ,elmwood, sandford, woodley, ojo, dorothy, ozma, scarecrow, miss, lord, tottenham, pumpkinhead, ...
	ground truth	miss milner you shall not leave the house this evening sir
	$w_k = 0$	miss millner you shall not leave the house this evening sir
Positive results	keywords	cap'n, booloroo, button, ghip, trot, pinkies , ghisizzle, blueskins, calder, bill, marianna, angareb, tiggie, ...
	ground truth	you are not like my people the pinkies and there is no place for you in our country
	$w_k = 0$	you are not like my people the pinkeys and there is no place for you in our country
Negative results	keywords	servius, praetors, senate, laws, solon, hovel, despotism, julian, decrees, athens , edicts, ...
	ground truth	the worthy friend of athanasius the worthy antagonist of ...
	$w_k = 0$	the worthy friend of athanasius the worthy antagonist of ..
	$w_k = 1.2$	the worthy friend of athenasius the worthy antagonist of ...

Table 4. Positive and negative samples of transcription on LibriSpeech with keyword boosting or not.

second row shows the result with boosting top 1% keywords and the last row with top 5% keywords. As we discussed in Section 3.3, the improvement in WER is relatively small. However in all scenarios, using the boosting method is effective in decreasing WER. Boosting with 5% of the keywords generally outperforms the other setups.

In addition, we can find that this method is more effective when the model is trained in a low-resource environment in Tables 2 and 3. This is an expected observation, because the model with not enough exposure in a target context is probably not familiar with the uncommon words. For a similar reason, the effect is larger without LM than with LM, since the LM should somewhat provide the context and keyword information. Furthermore, high w_k performs better in low-resource environment and moderate w_k is better in the higher-resource environment.

Table 3 shows the metrics mentioned in Section 3.3 on LibriSpeech test-clean. Although the precision is slightly decreased, the gain of recall is significant, resulting in a large gain in the F1-score. The same trend can also be observed in the other test sets. The keyword recall not only depends on w_k , but also on w_{LM} although w_k is the more influential factor. Our method is effective on every dataset, but in particular, on test-clean set of LibriSpeech we get a significant improvement of keyword recall from 94.5% to 98.5%. The LM often has negative effects on keyword recall in some datasets but it is consistently helpful to improving WER in every test scenario, as shown in Table 2.

In-house dataset. Table 5 shows the character error rate (CER) and the precision-recall values of the experiments done on our in-house datasets. We select larger values of w_k since this is found to be more effective for the Korean dataset. Even though the keywords do not make up a large proportion of the total word occurrences, there is a huge improvement in keyword recall especially from 61.9% to 82.3% in the *NAVER VLive* dataset, and a small but consistent improvement in CER. This proves that this method can be effectively utilised for real-world ASR services when the context can be specified.

3.5. Analysis of side-effects

Boosting a larger number of keywords than necessary causes *overboost* which makes the results worse for certain types of data. The reduction of precision in Table 3 show that some results recognise keywords more often than the actual number of occurrence, though this is still relatively rare. We can see an example of this in the last row of Table 4. The keyword **athens** is very similar to the word in ground truth **athanasius** so that **athenasius** has been recognised. Subtractive cost did not work as expected in this example, because the candidates containing the correct

data	w_k	CER	Precision	Recall	F1-score
Clova Note	0	8.07	98.9	91.9	95.3
	1	7.92	98.7	93.8	96.2
	3	7.81	98.4	95.5	96.9
	5	7.78	98.1	96.3	97.1
	7	7.90	97.7	96.5	97.1
NAVER VLive	0	16.74	95.8	61.9	75.2
	1	16.70	95.6	65.8	78.0
	3	16.62	93.8	74.0	82.7
	5	16.58	91.7	79.9	85.4
	7	16.60	87.5	82.3	84.8

Table 5. Result on *Clova Note* and *VLive* datasets

word had been pruned in the beam search due to its relatively lower scores than other boosted candidates. This is one of the corner cases that our method was not able to handle.

4. CONCLUSIONS

We proposed a new keyword-boosted beam search algorithm in speech recognition and demonstrated its performance with keywords extracted from the same book in the LibriSpeech dataset and with human-defined keywords in the in-house dataset. The result shows that this method is clearly helpful for recognising uncommon keywords that are important for understanding the context. We can use this method whenever we lack in-domain training dataset containing difficult keywords, but we only have the list of these words. It does not need any further text data or training process.

We observed that language models are helpful for improving keyword recall, but our boosting method is far more effective than LM at maintaining lower WER while mitigating the risk of overboost.

Our method has one requirement: we should have the keyword list in advance. We used TF-IDF on the training data of LibriSpeech to extract book-wise keywords, but there are many ways to obtain keywords in advance, such as using the list of characters in a TV show or the list of terminologies in a lecture. Strategies to effectively extract keywords are potential areas for future research.

5. ACKNOWLEDGEMENTS

We would like to thank Chan Kyu Lee and Icksang Han for their helpful advice.

6. REFERENCES

- [1] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., “Conformer: Convolution-augmented transformer for speech recognition,” *arXiv preprint arXiv:2005.08100*, 2020.
- [2] Wei Han, Zhengdong Zhang, Yu Zhang, Jiahui Yu, Chung-Cheng Chiu, James Qin, Anmol Gulati, Ruoming Pang, and Yonghui Wu, “Contextnet: Improving convolutional neural networks for automatic speech recognition with global context,” *arXiv preprint arXiv:2005.03191*, 2020.
- [3] Yu Zhang, James Qin, Daniel S Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V Le, and Yonghui Wu, “Pushing the limits of semi-supervised learning for automatic speech recognition,” *arXiv preprint arXiv:2010.10504*, 2020.
- [4] William Chan, Daniel Park, Chris Lee, Yu Zhang, Quoc Le, and Mohammad Norouzi, “Speechstew: Simply mix all available speech recognition data to train one large neural network,” *arXiv preprint arXiv:2104.02133*, 2021.
- [5] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *arXiv preprint arXiv:2006.11477*, 2020.
- [6] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, “SpecAugment: A simple data augmentation method for automatic speech recognition,” *arXiv preprint arXiv:1904.08779*, 2019.
- [7] Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huihui Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, “On using monolingual corpora in neural machine translation,” *arXiv preprint arXiv:1503.03535*, 2015.
- [8] Changhao Shan, Chao Weng, Guangsen Wang, Dan Su, Min Luo, Dong Yu, and Lei Xie, “Component fusion: Learning replaceable language model component for end-to-end speech recognition system,” in *Proc. ICASSP. IEEE*, 2019, pp. 5361–5635.
- [9] Felix Stahlberg, James Cross, and Veselin Stoyanov, “Simple fusion: Return of the language model,” *arXiv preprint arXiv:1809.00125*, 2018.
- [10] Pundak Golan, Sainath Tara N., Prabhavalkar Rohit, Kannan Anjuli, and Zhao Ding, “Deep context: End-to-end contextual speech recognition,” *IEEE Spoken Language Technology Workshop*, 2018.
- [11] Alon Uri, Pundak Golan, and Sainath Tara N., “Contextual speech recognition with difficult negative training examples,” *arXiv preprint arXiv:1810.12170*, 2018.
- [12] Jain Mahaveer, Keren Gil, Mahadeokar Jay, Zweig Geoffrey, Metzger Florian, and Saraf Yatharth, “Contextual rnn-t for open domain asr,” *arXiv preprint arXiv:2006.03411*, 2020.
- [13] Assaf Hurwitz Michaely, Mohammadreza Ghodsi, Zelin Wu, Justin Scheiner, and Petar Aleksic, “Unsupervised context learning for speech recognition,” in *IEEE Spoken Language Technology Workshop. IEEE*, 2016, pp. 447–453.
- [14] Petar Aleksic, Mohammadreza Ghodsi, Assaf Michaely, Cyril Allauzen, Keith Hall, Brian Roark, David Rybach, and Pedro Moreno, “Bringing contextual information to google speech recognition,” 2015.
- [15] Le Duc, Keren Gil, Chan Julian, Mahadeokar Jay, Fuegen Christian, and Seltzer Michael L., “Deep shallow fusion for rnn-t personalization,” *arXiv preprint arXiv:2006.03411*, 2020.
- [16] Young Mo Kang and Yingbo Zhou, “Fast and robust unsupervised contextual biasing for speech recognition,” *arXiv preprint arXiv:2005.01677*, 2020.
- [17] Ding Zhao, Tara N Sainath, David Rybach, Pat Rondon, Deepti Bhatia, Bo Li, and Ruoming Pang, “Shallow-fusion end-to-end contextual biasing,” in *Proc. Interspeech*, 2019, pp. 1418–1422.
- [18] Le Duc, Jain Mahaveer, Keren Gil, Kim Suyoun, Shi Yangyang, Mahadeokar Jay, Chan Julian, Shanguan Yuan, Fuegen Christian, Kalinli Ozlem, Saraf Yatharth, and Seltzer Michael L., “Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion,” *Proc. Interspeech*, 2021.
- [19] Graves Alex, “Sequence transduction with recurrent neural networks,” *ICML workshop on representation learning*, 2012.
- [20] Stephen Merity, Nitish Shirish Keskar, and Richard Socher, “Regularizing and Optimizing LSTM Language Models,” *arXiv preprint arXiv:1708.02182*, 2017.
- [21] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proc. ICML*, 2006, pp. 369–376.
- [22] “<https://librivox.org/>,” .
- [23] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *Proc. ICASSP. IEEE*, 2015, pp. 5206–5210.
- [24] “<https://www.openslr.org/11/>,” .
- [25] Takaaki Hori, Jaejin Cho, and Shinji Watanabe, “End-to-end speech recognition with word-based rnn language models,” in *IEEE Spoken Language Technology Workshop. IEEE*, 2018, pp. 389–396.
- [26] Papineni Kishore, “Why inverse document frequency,” *Proceedings of the North American Association for Computational Linguistic*, 2001.