Interspeech 2018
2-6 September 2018, Hyderabad, India

Study of Semi-supervised Approaches to Improving English-Mandarin Code-Switching Speech Recognition

Pengcheng Guo[1,2], *Haihua Xu*[2], *Lei Xie*[1,*], *Eng Siong Chng*[2,3]

[1] School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2] Temasek Laboratories, Nanyang Technological University, Singapore
[3] School of Computer Science and Engineering, Nanyang Technological University, Singapore

guopengcheng1220@gmail.com, lxie@nwpu.edu.cn, {haihuaxu,aseschng}@ntu.edu.sg

Abstract

In this paper, we present our efforts to improve the performance of a code-switching speech recognition system using semi-supervised training methods from lexicon learning to acoustic modeling, on the South East Asian Mandarin-English (SEAME) data. We first investigate semi-supervised lexicon learning approach to adapt the canonical lexicon, which is meant to alleviate the heavily accented pronunciation issue within the code-switching conversation of the local area. As a result, the learned lexicon yields improved performance. Furthermore, we attempt to use semi-supervised training to deal with those transcriptions that are highly mismatched between the human transcribers and the ASR system. Specifically, we conduct semi-supervised training assuming those poorly transcribed data as unsupervised data. We found the semi-supervised acoustic modeling can lead to improved results. Finally, to make up for the limitation of the conventional n-gram language models due to the data sparsity issue, we perform lattice rescoring using neural network language models, and significant WER reduction is obtained.

Index Terms: speech recognition, code-switching, lexicon learning, semi-supervised training, lattice rescoring

1. Introduction

In recent years, code-switching speech recognition research has been drawn increasingly attention in speech recognition community. This is because most existing state-of-the-art speech recognition engines are only capable of understanding a specific monolingual language, say English or Mandarin. When people converse with mixed language simultaneously, for instance, "请问 到 changi airport 怎么 走?", monolingual speech recognition engines failed for the code-switching part. This is common in South East Asia area, where people can usually speak several languages and code-switching conversation is very common [1].

Compared with the building of a monolingual speech recognition system, it is much more challenging to build a code-switching speech recognition system, since one is often encountered with data sparsity problem for either individual languages.

First, to address the data sparsity issue in acoustic modeling, cross-lingual phone merging [2, 3, 4] and global phone set usage [5] was widely studied in the era of GMM-HMM acoustic modeling. This indeed enforced cross-lingual data sharing and alleviated the intensity of individual language data sparsity issue to some degree, but it also brought about cross-lingual confusion problem. That is, inter-language substitution issue arises due to the discriminative capability of the merged phone being undermined. The underlying reason is that there are really some phones that are acoustically similar but their acoustic contexts are significantly different in each individual languages. However, with the advent of deep neural network (DNN) acoustic modeling techniques, the necessity of cross-lingual phone merging or global phone usage is decreased thanks to the topology of DNN architecture, since the lower hidden layers are inherently shared among different languages and can be jointly learned by each individual languages [6, 7].

Secondly, it is also challenging to robustly estimate language models (LMs) for code-switching speech recognition. Take the conventional bigram LMs as an example. It is easy to access sufficient counts for the bigrams within each language; however, it is hard to get those cross-lingual bigram counts, such as "到 changi" or "airport 怎么" etc. as is seen from the above exemplary utterance. On the one hand, we can employ machine translation method to generate a lot of artificial code-switching text utterances [8]. However limited performance improvement was obtained due to a lot of noisy cross-lingual n-gram counts being introduced. On the other hand, we can use neural networks based language modeling method, which learns n-gram probability in continuous space and is more powerful to generalize [9, 10, 11].

In this paper, we put efforts into improving code-switching speech recognition with semi-supervised training approaches, assuming the imperfection of the original lexicon and human acoustic transcriptions.

For the lexicon problem, since there is no off-the-shelf South East Asian English lexicon available, we use the CMU open source lexicon to start with. Obviously, there are a pronunciation mismatch problem and an out-of-vocabulary (OOV) problem as well. We first collect alternative pronunciations from original lexicon, G2P and phonetic decoding. Then we adopt a data likelihood based criterion to select and prune all pronunciation candidates.

Once the lexicon is learned, we fix it and analyze the transcription quality. Word matched error rate (WMER) is computed to score the human transcriptions and the ASR decode hypotheses as in [12]. We found the WMER is around 10% on the overall training data and there is about 7 hours of data with WMER over 30%. To deal with these data, we propose to use lattice-free maximum mutual information (LF-MMI) based semi-supervised training method. We found that, with semi-supervised training, the poorly transcribed data can also improve the system performance.

This paper is organized as follows: Section 2 describes the distribution of the overall SEAME data. Section 3 reports the experimental setup. Section 4 proposes our approach to semi-supervised lexicon learning. Section 5 is for semi-supervised

∗ Corresponding author.

10.21437/Interspeech.2018-1974

acoustic modeling. Section 6 presents neural network language model based lattice rescoring and we conclude in Section 7.

2. Data

2.1. Overall description

The SEAME corpus is a microphone based spontaneous conversational bilingual speech corpus, in which most of the utterances contain both English and Mandarin uttered by 154 oversea Chinese in Malaysia and Singapore areas [8, 13]. Figure 1 describes the distribution of speakers with regard to Mandarin occurrence rate per speaker. We can see the SEAME corpus is generally biased with Mandarin.

Specifically, about 90% speakers speak code-switching utterances with Mandarin rate ranged from 10% to 90%, while about 2% speakers have utterances with Mandarin rates no more than 10%, and 5% speakers have utterances over 90% being Mandarin. Besides, we find Singaporean speakers normally have more English words in their utterances, while Malaysian speakers are more likely to converse with utterances dominated by Mandarin.

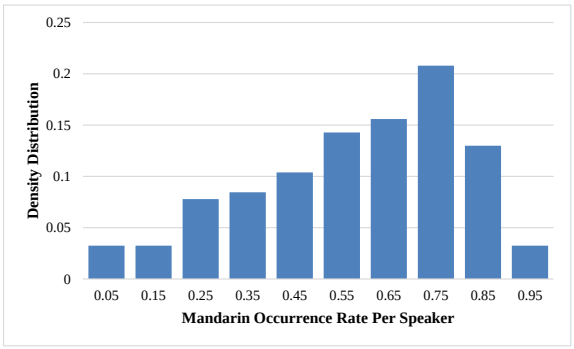

Figure 1: *Distribution of speakers versus Mandarin occurrence rates per speaker for the overall SEAME data. Each bar's statistics come from the* ± 0.05 *range of the corresponding data on the horizontal axis.*

2.2. Test data definition

To evaluate the proposed methods, we define two evaluation data sets and each is randomly selected from 10 gender balanced speakers. However, they are defined differently, and one is dominated by Mandarin, named as $eval_{man}$, and the other is dominated by English, as $eval_{sge}$. The detailed statistics are revealed in Table 1. We think these "biased" data sets would give more clues to show the effectiveness of each proposed methods on each individual languages.

Table 1: *The statistics of two evaluation data sets, one is dominated by Mandarin, named* $eval_{man}$*, another is dominated by English, named* $eval_{sge}$*.*

	Train	$eval_{man}$	$eval_{sge}$
Speakers	134	10	10
Duration (hrs)	101.1	7.5	4
Mandarin (%)	0.59	0.69	0.29

3. Experimental setup

3.1. Acoustic modeling

We use time-delay neural networks (TDNNs) [14, 15] trained with lattice-free MMI (LF-MMI) [16, 17] to build acoustic models in all experiments. The front-ends are made up of 40-dimensional MFCC plus 100-dimensional i-vectors [18]. They are LDA transformed over ± 2 concatenated MFCC features plus i-vectors before fed to the TDNN. For training, the supervisions obtained from the senones of the prior trained GMM-HMM acoustic models.

3.2. Lexicon and language modeling

The original lexicon is composed of a CMU English lexicon [19] and a Mandarin lexicon [20]. Since we do not merge cross-language phonemes, we separate the two phoneme sets by adding a suffix string of language identity for each phoneme to differentiate in our lexicon. Overall, we have 252 phonemes, of which there are 213 Mandarin and 39 English phonemes respectively. For language modeling, only the transcriptions of the training part of the SEAME data are employed.

3.3. Baseline results

We report the baseline results on the two evaluation sets as defined in Table 1. To prepare the data, we normalize the overall data as follows: 1) remove those training data that contains $\langle unk \rangle$ words; 2) manually correct about 400 obvious word typos from the human transcriptions. We then train the TDNN acoustic models using the LF-MMI frame subsampling method, achieving the WER of 23.39% on $eval_{man}$ and 33.04% on $eval_{sge}$ respectively. All experiments are based on the Kaldi toolkit [21].

4. Semi-supervised lexicon learning

4.1. Motivation

As mentioned, our available lexicon is composed of the standard American English and Chinese lexicons, which are mismatched with the SEAME acoustic data, particularly for the English pronunciation part. The mismatched problems are due to two reasons. One is that the large amount of local spoken words leads to a high OOV rate (1.10%) on the SEAME data. Another reason is that the English pronunciation of the SEAME data is different from the American English. For instance, word "three" is normally pronounced as "tree" in the SEAME data. In this section, we propose to use the supervised G2P training and the unsupervised phonetic decoding to address the mismatched problems, which named semi-supervised lexicon learning.

4.2. Method

We first address the OOV problem using the G2P method [22]. As a result, we obtain an OOV-free lexicon though pronunciation mismatched problem persists. To solve the pronunciation mismatched problem, we employ a phonetic level decoding process on the training data as inspired in [23]. We try to learn adapted word pronunciations using the time boundary of the corresponding word and decoded phone sequence. As this unsupervised procedure would introduce noisy pronunciations for some words, pronunciation probability estimation and pruning are crucial. We adopt the estimation and pruning approaches proposed in [24], and the whole procedure of the method is depicted in Figure 2.

1: **Input**: Original lexicon as L_0
2: Train G2P model using L_0 and generate lexicon L_1 for OOV words
3: Merge lexicons L_0 and L_1 to train GMM-HMM models
4: Use phonetic LMs to decode the training data
5: Use word and phone time boundaries to generate adapted lexicon L_2 from decoded data
6: Merge lexicons L_0, L_1, L_2 as lexicon L'
7: Use lexicon L' to generate lattices for each training utterances
8: Estimate the pronunciation probability $p(w, b)$, which means the probability of word w be pronounced as b, to maximize the data likelihood $p(O_u, w, b)$, which means the probability of word w be pronounced as b in utterance O_u
9: Compute the reduction of data likelihood before and after removing a specific pronunciation b, and prune the least reduction pronunciation
10: **Output**: Learned lexicon L with pronunciation probability

Figure 2: *The procedure of the semi-supervised lexicon learning for the SEAME code-switching speech recognition*

4.3. Results

Table 2 reports various WER results on the two evaluation sets with different lexicons. The baseline results are obtained with the original lexicon that has the OOV rate about 1.10%, while the remaining lexicons have no OOV word. In Table 2, we notice that the G2P method is very effective and makes the WER drop obviously. Furthermore, the semi-supervised lexicon learning is beneficial, and further WER drops are gained over the G2P lexicons consistently.

Table 2: *WER (%) results on both $eval_{man}$ and $eval_{sge}$ evaluation sets with different lexicons, the "#pron/word" stands for the statistics of average pronunciations per word.*

Lexicon	#pron/word	$eval_{man}$	$eval_{sge}$
Original Lex	1.11	23.39	33.04
G2P 1-best	1.12	22.85	32.12
G2P 5-best	1.31	22.82	31.99
Learned Lex	1.39	22.76	31.89

5. Semi-supervised acoustic modeling

5.1. Motivation

Transcribing code-switching data is challenging as it requires our transcribers to have a strong multilingual background and follow the ASR orthographic transcription protocol. Therefore, we are assuming there are some errors in human transcriptions. To verify our assumptions, we use our best ASR system to decode the training data and compare the decoded hypotheses with the manual transcriptions. We use the WMER to measure the quality of the human transcriptions.We find the WMER is around 10% on the overall training data, and there are about 7 hours of data whose WMER is over 30%. Table 3 reports the cumulative distribution of the WMER on the overall training data.

In order to reduce the bad effect of those "poorly" transcribed data, we can simply remove such a part of data during acoustic model training. However, if the data with higher WMER can be exploited, better acoustic modeling might be yielded. From this standpoint, we propose to use semi-supervised training, regarding those higher WMER data as unsupervised data.

Table 3: *Cumulative distribution of the selected unsupervised data according to threshold WMER (%)*

WMER (%)	Durations (hrs)	Rate (%)
>0	67.18	66.43
>20	14.21	14.05
>30	6.99	6.91
>40	3.96	3.92

5.2. LF-MMI based Semi-supervised training

To fully exploit those acoustic data with higher WMER, we use the LF-MMI training method to conduct semi-supervised training as advocated in [25]. Unlike conventional semi-supervised training methods, which use frame-level, word-level, or utterance-level confidence scores [26, 27] to select supervisions, the proposed method in [25] uses the whole lattice as supervision.

We perform semi-supervised training as the following steps. First, we use the best LF-MMI acoustic models to decode those "unsupervised" data, obtaining lattices of alternative pronunciations. Then, the word lattices are converted to phone lattices as in [17]. Finally, we make senone supervisions from the phone lattices and do MMI training by weighting the lattice supervisions with an LM scale and the posterior of the best path from the decoded lattices.

5.3. Results

Table 4 reports the WER results of the two semi-supervised LF-MMI training methods, one using the best path from the lattice as supervision, and the other using the pruned lattice as supervision. The baseline results are the best results in Table 2.

In Table 4, several points are useful. First, when we remove the data with WMER over 30%, we can consistently obtain improved results compared with the baseline result where the whole training data is used for acoustic model training. This confirms our assumption that the human transcriptions of the SEAME data may have some errors. Secondly, LF-MMI based semi-supervised training is useful. It makes an improvement over both the baseline system and the system trained with fewer data by means of the removal of the data containing errors. Lastly and more importantly, though poorly transcribed data can be detrimental to the LF-MMI training (from the "Baseline" in Table 4), we still can exploit the data to contribute the final system by means of semi-supervised training (from the last two rows of Table 4).

Moreover, to investigate the overall transcription quality of the SEAME data, and the contribution of the LF-MMI semi-supervised training method, we plot Figures 3 and 4 of the WER results versus the different portion of data used as unsupervised data, respectively. In both figures, we can see the majority of the SEAME transcriptions are reliable. Either the "removal method" or "lattice supervision" based semi-supervised training yields degraded results if we use the data whose WMER is over 20% (about 14 hours of data) as poorly transcribed data. However, when we use the data with WMER over 30% (about

Table 4: *The WER (%) results of the semi-supervised training methods with different supervisions, where we use the acoustic data whose WMER* $> 30\%$ *as unsupervised data.*

System	$eval_{man}$	$eval_{sge}$
Baseline	22.76	31.89
Removal method	22.69	31.66
Best path supervision	22.63	31.78
Lattice supervision	22.57	31.59

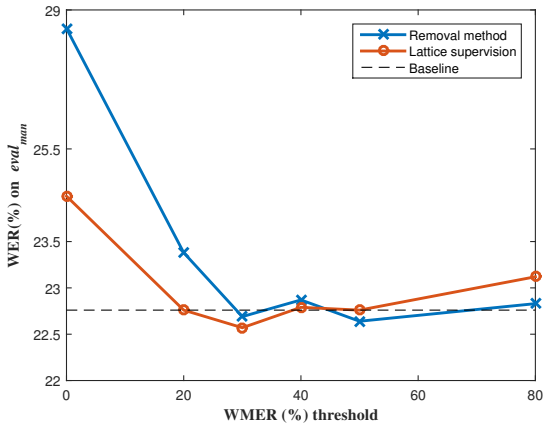

Figure 3: *WER (%) results of the lattice supervision based LF-MMI semi-supervised training on* $eval_{man}$ *set, using the acoustic data whose WMER (%) is over the specified threshold as unsupervised data.*

7 hours of data), both two methods make improved results compared with the baseline system. These mean the poorly transcribed data lies in $(7, 14)$ hours.

6. Lattice rescoring

Recently, it has been widely demonstrated the effectiveness of the recurrent neural network language models (RNNLMs) [28] over the conventional n-gram LMs in terms of perplexity and rescoring results. The main advantages of the RNNLMs lie in two aspects. First, RNNLMs can exploit much longer word history compared with the n-gram LMs. Secondly, it estimates word probability in continuous space, alleviating data sparsity issue. However, the drawback of the RNNLMs is that it is hard to be employed in online decoding due to high computational cost. Hence they are usually used in N-best or lattice rescoring.

In this paper, since we only use the training transcription to train 4-gram LMs, it has a data sparsity issue. Therefore, it is worthwhile to use RNNLMs as a comparison. In practice, we use the Kaldi RNNLM toolkit [29] to train RNNLMs and rescore the lattice. Table 5 reports our lattice rescoring results. In Table 5, it can be seen that the lattice rescoring method achieves the best result with a WER of 20.54% on the $eval_{man}$ data and a WER of 29.56% on the $eval_{sge}$ data. It achieves 9% and 8.8% relative WER reductions respectively over the baseline system obtained with the best semi-supervised training configuration.

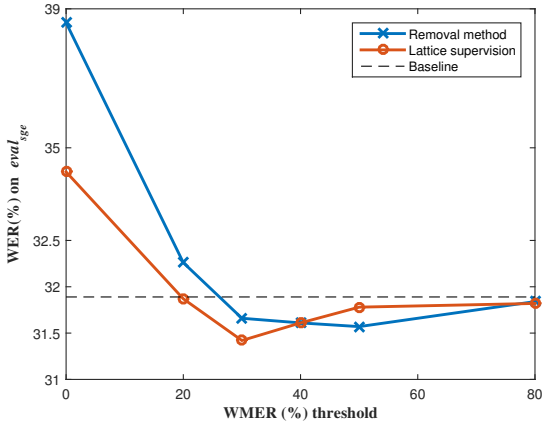

Figure 4: *WER(%) results of the lattice supervision based LF-MMI semi-supervised training on* $eval_{sge}$ *set, using the acoustic data whose WMER (%) is over the specified threshold as unsupervised data.*

Table 5: *WER(%) results using lattice rescoring*

System	$eval_{man}$	$eval_{sge}$
Best 4-gram	22.57	32.42
N-best RNNLM rescoring	21.14	30.15
Lattice RNNLM rescoring	20.54	29.56

7. Conclusions

In this paper, we studied semi-supervised training approaches to lexicon learning and acoustic modeling under code-switching speech recognition scenarios. The semi-supervised lexicon learning is to deal with the pronunciation mismatch problems. By means of collecting word pronunciation candidates and pruning them, we gained a learned lexicon that is more appropriate for the target SEAME data. As a result, we obtained better recognition results with the learned lexicon. Furthermore, we also proposed to use LF-MMI semi-supervised training to deal with the uncertainties of human transcriptions. That is, we regard the data whose transcriptions have higher WMER as unsupervised data. We also gained improved recognition results from the proposed semi-supervised acoustic modeling. It should be noted that there may exist similar transcribed errors in the evaluation data, and we will take it into account in the future work. Lastly, to alleviate the data sparsity issue for language modeling, we employed recurrent neural network language modeling method to rescore lattice, and achieved significant word error rate reduction.

8. Acknowledgments

The research work is supported by the National Key Research and Development Program of China (Grant No.2017YFB1002102) and the National Natural Science Foundation of China (Grant No.61571363).

9. References

[1] P. Auer, *Code-switching in conversation: Language, interaction and identity*. Routledge, 2013.

[2] B. Mak and E. Barnard, "Phone clustering using the bhattacharyya distance," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 4. IEEE, 1996, pp. 2005–2008.

[3] Y. Li, P. Fung, P. Xu, and Y. Liu, "Asymmetric acoustic modeling of mixed language speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5004–5007.

[4] H. Lin, L. Deng, D. Yu, Y.-f. Gong, A. Acero, and C.-H. Lee, "A study on multilingual acoustic modeling for large vocabulary asr," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4333–4336.

[5] IPA, *Handbook of the International Phonetic Association : a guide to the use of the International Phonetic Alphabet*. Cambridge University Pres, 1999.

[6] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7304–7308.

[7] E. Yılmaz, H. van den Heuvel, and D. van Leeuwen, "Investigating bilingual deep neural networks for automatic recognition of code-switching frisian speech," *Procedia Computer Science*, vol. 81, pp. 159–166, 2016.

[8] N. T. Vu, D.-C. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E.-S. Chng, T. Schultz, and H. Li, "A first speech recognition system for mandarin-english code-switch conversational speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4889–4892.

[9] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[10] H. Adel, N. T. Vu, F. Kraus, T. Schlippe, H. Li, and T. Schultz, "Recurrent neural network language modeling for code switching conversational speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8411–8415.

[11] H. Adel, N. T. Vu, and T. Schultz, "Combination of recurrent neural networks and factored language models for code-switching language modeling," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, 2013, pp. 206–211.

[12] P. Bell, M. J. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester *et al.*, "The mgb challenge: Evaluating multi-genre broadcast media recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 687–693.

[13] D.-C. Lyu, T.-P. Tan, E.-S. Chng, and H. Li, "An analysis of a mandarin-english code-switching speech corpus: Seame," *Age*, vol. 21, pp. 25–8, 2010.

[14] A. Waibel, "Modular construction of time-delay neural networks for speech recognition," *Neural computation*, vol. 1, no. 1, pp. 39–46, 1989.

[15] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[16] D. Povey, M. Hannemann, G. Boulianne, L. Burget, A. Ghoshal, M. Janda, M. Karafiát, S. Kombrink, P. Motlíček, Y. Qian *et al.*, "Generating exact lattices in the wfst framework," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4213–4216.

[17] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi." in *Interspeech*, 2016, pp. 2751–2755.

[18] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors." in *ASRU*, 2013, pp. 55–59.

[19] "Cmu pronunciation dictionary for american english," http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[20] R. Hsiao, M. Fuhs, Y.-C. Tam, Q. Jin, and T. Schultz, "The cmu-interact 2008 mandarin transcription system," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[22] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech communication*, vol. 50, no. 5, pp. 434–451, 2008.

[23] A. Laurent, S. Meignier, T. Merlin, and P. Deléglise, "Acoustics-based phonetic transcription method for proper nouns," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[24] X. Zhang, V. Manohar, D. Povey, and S. Khudanpur, "Acoustic data-driven lexicon learning based on a greedy pronunciation selection framework," *arXiv preprint arXiv:1706.03747*, 2017.

[25] V. Manohar, H. Hadian, D. Povey, and S. Khudanpur, "Semi-supervised training of acoustic models using lattice-free mmi."

[26] K. Vesely, M. Hannemann, and L. Burget, "Semi-supervised training of deep neural networks," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 267–272.

[27] S. Thomas, M. L. Seltzer, K. Church, and H. Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6704–6708.

[28] T. Mikolov and G. Zweig, "Context dependent recurrent neural network language model." *SLT*, vol. 12, pp. 234–239, 2012.

[29] H. Xu, K. Li, Y. Wang, J. Wang, S. Kang, X. Chen, D. Povey, and S. Khudanpur, "Neural network language modeling with letter-based features and importance sampling," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on. IEEE*, 2018.