INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

On the End-to-End Solution to Mandarin-English Code-switching Speech Recognition

Zhiping Zeng[1], *Yerbolat Khassanov*[2], *Van Tung Pham*[1], *Haihua Xu*[1], *Eng Siong Chng*[1,2], *Haizhou Li*[3]

[1]Temasek Laboratories, Nanyang Technological University, Singapore
[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[3]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
zengzp@ntu.edu.sg

Abstract

Code-switching (CS) refers to a linguistic phenomenon where a speaker uses different languages in an utterance or between alternating utterances. In this work, we study end-to-end (E2E) approaches to the Mandarin-English code-switching speech recognition task. We first examine the effectiveness of using data augmentation and byte-pair encoding (BPE) subword units. More importantly, we propose a multitask learning recipe, where a language identification task is explicitly learned in addition to the E2E speech recognition task. Furthermore, we introduce an efficient word vocabulary expansion method for language modeling to alleviate data sparsity issues under the code-switching scenario. Experimental results on the SEAME data, a Mandarin-English code-switching corpus, demonstrate the effectiveness of the proposed methods.

Index Terms: Code-switching, speech recognition, end-to-end, multitask learning, language identification

1. Introduction

Code-switching (CS) is a linguistic phenomenon, where speaker's utterances contain different languages, either inside a given utterance or between utterances. It frequently appears in world wide areas. Therefore, developing a code-switching speech recognition (CSSR) system is important and has received increasing attention recently.

While DNN-HMM-based automatic speech recognition (ASR) framework is popular in code-switching speech recognition [1, 2], it has some clear limitations. Firstly, one needs to build a big lexicon mixed with words from different languages, and it would take more human efforts to label pronunciations for those words from different languages. Secondly, acoustic modeling (AM), language modeling (LM) and lexicon modeling components of the DNN-HMM-based ASR system, which are optimized separately. This would lead to sub-optimal performance.

In this paper, we pursue an End-to-End (E2E) strategy to resolve Mandarin-English code-switching speech recognition instead. In contrast to the DNN-HMM-based approach, it doesn't require any lexicon modeling efforts. More importantly, the entire recognition system comprises compactly connected neural networks that are jointly learned from scratch. To our best knowledge, this is the first attempt of using E2E strategy to code-switching speech recognition task. Our contributions mainly lies in the following aspects.

Firstly, we manage to build competitive E2E baseline systems using data augmentation and byte-pair encoding (BPE) based subword units [3–5]. We found data augmentation is more effective to the E2E framework than the DNN-HMM-based one. Besides, we found the BPE subword units yield better recognition results than characters. This is consistent with what is reported in [3].

Secondly, we employ a multitask learning (MTL) [6] method to enhance our E2E-based code-switching ASR system. Specifically, we propose to use language identification (LID) as the auxiliary task to help improving the speech recognition performance. It showed that the LID-based MTL helps on Token Error Rate (TER) reduction.

Thirdly, to alleviate the cross-lingual data sparsity issue in language modeling, we introduce a word vocabulary expansion method inspired by [7]. Note that in [7], this technique is applied for monolingual data to improve the speech recognition performance when its output is rescored by language modeling, while this work applies it to the code-switching language model rescoring.

The paper is organized as follows. Related work are presented in Section 2. Then, proposed multitask learning E2E code-switching speech recognition approaches and code-switching word vocabulary expansion LM rescoring are introduced in Section 3. Experimental setups are described in Section 4, then experimental results and analysis are reported in Section 5. Finally, we conclude and talk about future work in Section 6.

2. Related work

Previous works on code-switching speech recognition relied on the conventional GMM-HMM or DNN-HMM framework [1, 2, 8, 9]. Recently, end-to-end speech recognition methods have drawn much attention and produced promising recognition results [10–17]. There are two main directions for end-to-end speech recognition. One is the earlier proposed Connectionist Temporal Classification (CTC) [10], and the other is the attention mechanism based method [11]. CTC performed well on various corpora such as WSJ [12, 13] and SWB [14]. Recently, inspired from the attention-based machine translation framework [15], attention-based E2E began to play a crucial role in speech recognition, achieving the state-of-the-art results [16]. Despite the two methods are different, researchers in [17] exploited the advantages of each methods to build the hybrid CTC/attention based E2E ASR system, which leads to better results compared to an ASR system built with either single method.

We apply end-to-end approach to code-switching speech recognition task in this paper. This differs from the previous end-to-end speech recognition that has worked only for monolingual case. Recently, it was shown that E2E method can be employed to perform multilingual speech recognition simulta-

http://dx.doi.org/10.21437/Interspeech.2019-1429

neously [18, 19]. However, the multilingual speech recognition is not a code-switching task.

3. Approaches to end-to-end CSSR

In this section, we present various approaches to achieve better end-to-end code-switching speech recognition. We first aim to build competitive baselines. To start with, we investigate data augmentation method and study different subword units as the output of the end-to-end CSSR system. After that, we employ a multitask learning by introducing the LID as the auxiliary task to boost the performance of our end-to-end CSSR. Finally, we propose a modified neural language modeling framework, aiming to alleviate the cross-lingual data sparsity issue within the CSSR task.

3.1. Approaches to develop E2E CSSR baselines

One major challenge of building the E2E system is that it requires a lot of data to train the model [20]. To deal with this problem, we apply speech speed perturbation based data augmentation method proposed in [21, 22], as the effectiveness of the method has been proved in the conventional DNN-HMM ASR method. By manipulation, we obtained x3 times of the original data, with a speaking rate of 90%, 100%, and 110% of the original data respectively. In this work, E2E with data augmentation is one of our baseline.

Another issue for E2E system is how to select the output units. As we are dealing with both Mandarin and English output units simultaneously, it sounds straightforward to use characters as the units. However, this will result in only 26 output units for English and several thousand for Mandarin. We conjecture that such an unbalanced situation will be disadvantageous to English, yielding worse recognition results. To balance the units between the two languages, we decide to use subword units [16] for English. BPE subword units have shown to be helpful for English [3]. As a result, in this work, we use the BPE subwords for English, while leaving the output units of Mandarin fixed with characters. Therefore, another baseline in this work is to apply BPE subword units on the E2E system trained with data augmentation.

3.2. Multitask Learning of CSSR with LID

Inspired from the work in [17], we adopted a hybrid CTC/attention based E2E architecture to conduct CSSR.

Specifically, let X be the input acoustic sequence, Y be the output sequence comprising characters or BPE units, $\mathcal{L}_{CTC}(Y|X)$ be the CTC objective loss, $\mathcal{L}_{att}(Y|X)$ be the attention-based objective loss. Then, the objective loss $\mathcal{L}_{MTL}(Y|X)$ of the entire hybrid E2E system is as follows:

$$\mathcal{L}_{MTL}(Y|X) = \lambda_1 \mathcal{L}_{att}(Y|X) + (1-\lambda_1)\mathcal{L}_{CTC}(Y|X) \quad (1)$$

where $\lambda_1 \in [0, 1]$ is a hyper-parameter to control the contribution of each model.

For the code-switching system, although we can infer the language identification from the decoding transcription, the language information has not been used explicitly during training. We believe that using the language identification would improve the CSSR performance. To this end, we extend the multitask learning method as indicated in Eq (1). The framework is illustrated in Figure 1. As a result, the whole training objective loss is changed as follows:

$$\begin{aligned}\mathcal{L}_{MTL}(Y|X) = \lambda_1 \mathcal{L}_{att}(Y|X) + (1-\lambda_1)\mathcal{L}_{CTC}(Y|X) \\ + \lambda_2 \mathcal{L}_{lid}(Z|X)\end{aligned} \quad (2)$$

where Z is the output LID sequence, $\mathcal{L}_{lid}(Z|X)$ represents the LID objective loss, and we restrict $\lambda_2 \in [0, 1]$.

As indicated in Figure 1, we have investigated two methods to incorporate the LID into the hybrid CTC/attention framework. One is to share the same attention model with the speech recognition task (LID_{shared}), and another is to learn an independent attention model by itself (LID_{indep}). Both methods use the same objective function as in the Eq (2).

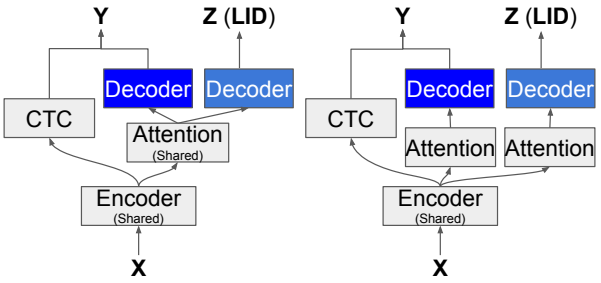

Figure 1: *A multitask learning (MTL) framework with language identification (LID) for E2E-based CSSR. The LID task can share the same attention module (left) or use a separate attention module (right).*

3.3. Vocabulary expansion for neural language model

The performance of E2E speech recognition can be further improved when its output is rescored by neural language model (NLM) [11]. However, the vocabulary coverage of the NLM is usually a shortened list of the entire ASR vocabulary, and such the list is usually only composed of most frequent words, due to the necessity of learning complexity restriction. Consequently, the probability of those infrequent words that are out-of-shortlist are not well learned by the NLM.

In code-switching speech recognition scenario, we treat those words that occur at the cross-lingual transition positions as 'infrequent' words, in addition to other 'infrequent' words that we have in monolingual text. The main idea is to 'borrow' probability mass for such 'infrequent' words (target words) from those words that are in the shortlist and semantically close to the target words, as advocated in [7].The benefit of using the method in [7] is that we don't have to learn a big NLM. However, the downside of the method is that word semantic clustering is needed beforehand. This can be done with word embedding vector. For the more details of implementation, one can refer to [7].

4. Experimental setup

4.1. Data

We conduct experiments on the SEAME corpus [23] which is a spontaneous conversational Mandarin-English code-switching speech corpus. The duration of utterances that contain code-switching is about 68% in our training set. We define two test sets that contains code-switching speech [2]. Detailed division of SEAME Corpus can be seen in Table 1. One is biased to

Mandarin speech (denoted as dev_{man}), and another is biased to Southeast Asian accent English (dev_{sge}). Each test set contains 10 speakers with balanced genders[1]. In what follows, we report token error rate (TER, Chinese character and English word respectively) on the test data.

Table 1: *The detailed division of SEAME Corpus,**'Man'**, **'En'** and **'CS'** mean pure Mandarin, pure English and Mandarin-English code-switch inside utterance.*

	Speakers	Hours	Duration Ratio		
			Man	En	CS
train	134	101.13	16%	16%	68%
dev_{man}	10	7.49	14%	7%	79%
dev_{sge}	10	3.93	6%	41%	53%

4.2. DNN-HMM baseline system

Besides two baselines in Section 3.1, we also build a DNN-HMM system as another baseline for comparison. We use Kaldi toolkit [24] to train a lattice-fee maximum multual information (MMI) based time delay neural network (TDNN) [25]. The TDNN has 6 hidden layers with 1024 hidden units, and the input features are 40 dimensional MFCC plus 100 dimensional i-vectors. The outputs are senones that are langauge independent. For language modeling, we only use the transcriptions of the training part of the SEAME data to train the 4-gram language model.

4.3. E2E ASR system setup

We use ESPnet toolkit[26][2] to train our E2E ASR system. The encoder consists of one-layer CNN and six-layers BLSTM with 320 hidden units. The decoder consists of one-layer LSTM with 320 hidden units. CTC weight $(1 - \lambda_1)$ is fixed with 0.2. The attention method used in this work is a combination of content-based and location-based methods [27]. To train a BPE subword model, setting character coverage rate for 0.9995 to determine the minimum Mandarin-English mixed symbols, which results in minimum 1806 character symbols. Since we attempt to build 4 BPE subword models with the vocabulary size 1.9k, 2k, 3k and 4k units, all that have dictionary bigger than 1806.

5. Experimental results and analysis

5.1. Results of the E2E CSSR system

Table 2 reports our TER results of the E2E CSSR system with different setups, using those from the Kaldi LF-MMI TDNN CSSR system as a contrast.

We have three observations from in Table 2. Firstly, our E2E ASR systems are not as competitive as the LF-MMI TDNN systems in general. This suggests that further effort is required to improve the E2E ASR system at least on the limited training data. Secondly, data augmentation significantly helps on TER reduction for our E2E ASR system, suggesting that more data might further reduce the performance gap between our E2E and LF-MMI TDNN systems. Thirdly, the BPE subword units are much more effective than the character units and the 3k BPE produces best results.

[1]Our KALDI format based test sets are released in the following link: https://github.com/zengzp0912/SEAME-dev-set

[2]https://github.com/espnet/espnet

Table 2: *The TER of different E2E CSSR systems as compared to the LF-MMI TDNN ASR counterparts.*

System	Data Aug	Subword	TER (%)	
			dev_{man}	dev_{sge}
Kaldi-TDNN	No	N.A	23.5	32.0
Kaldi-TDNN-DA	Yes	N.A	**22.1**	**30.9**
E2E-CHAR	No	Character	34.5	46.4
E2E-DA-CHAR	Yes	Character	26.5	38.4
E2E-DA-BPE1.9k	Yes	BPE	26.7	36.3
E2E-DA-BPE2k	Yes	BPE	26.6	35.9
E2E-DA-BPE3k	Yes	BPE	26.4	36.1
E2E-DA-BPE4k	Yes	BPE	26.6	36.2

One of the differences between the CSSR and the monolingual ASR is that there is cross-lingual substitutions for the CSSR. Table 3 reports various kinds of substitution rates from different E2E systems. We note from Table 3 that there are much fewer cross-lingual substitution than same language substitutions, and they are about 10% on each test sets. Besides, it can be seen that $S_{E\to M}$ is higher than $S_{M\to E}$ by $\sim$5% which indicates that more English words is substituted by Mandarin characters than the other way.

Table 3: *Different substitution (Sub) errors, where **'M'** and **'E'** stand for Mandarin and English respectively.*

Sub	System	dev_{man} (%)	dev_{sge} (%)
$S_{E\to E}$	E2E-DA-CHAR	30.8	38.5
	E2E-DA-BPE-3k	20.9	25.3
$S_{E\to M}$	E2E-DA-CHAR	8.4	5.0
	E2E-DA-BPE-3k	8.0	5.2
$S_{M\to E}$	E2E-DA-CHAR	2.7	5.2
	E2E-DA-BPE-3k	3.0	5.1
$S_{M\to M}$	E2E-DA-CHAR	12.9	14.7
	E2E-DA-BPE-3k	13.0	14.7

Table 4 shows the top 5 examples of cross-lingual substitution on both test sets. From Table 4, most of the cross-lingual substitution are only involved with those non-content or acoustically similar modal words that are rather short.

Table 4: *Top 5 cross-lingual substitution examples, where **'M'** and **'E'** stand for Mandarin and English respectively.*

$S_{M\to E}$			$S_{M\to E}$		
Ref	Hyps	count	Ref	Hyps	count
ah	啊	87	的	**the**	39
eh	诶	75	咯	lor	36
er	呃	38	哦	oh	21
the	的	37	有	**you**	18
oh	哦	36	诶	eh	17

5.2. Effect of MTL with LID

Figure 2 shows the TER of the two MTL with LID for E2E CSSR methods, LID_{shared} and LID_{indep}, at different weighting factor λ_2 in Eq (2). The baseline results are the corresponding best results from Table 2, i.e. E2E-DA-CHAR and E2E-DA-BPE3k. We observe that the two methods improve the results in most cases, and both methods can yield the best results when

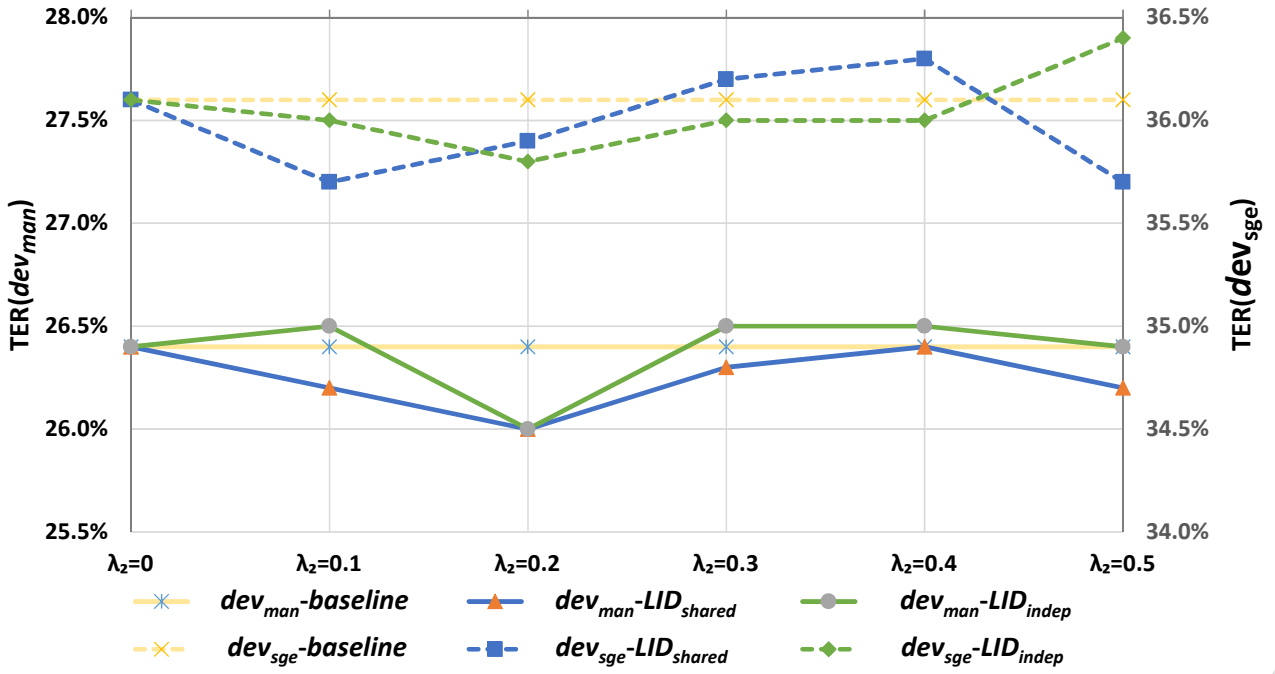

Figure 2: *TER results of the MTL with LID for E2E CSSR: TER versus LID weighting factor (λ_2), dev_{man} is indicated by left axis, while dev_{sge} is indicated by right axis.*

λ_2 is around 0.2. However, there are no obvious difference between the two methods.

Table 5 reveals the TER results of the two E2E MTL framework with LID CSSR methods with $\lambda_2 = 0.2$. Overall, the proposed methods yield improved results though not significant. It seems more effective when it is applied to the character based system.

Table 5: *TER(%) results of the two E2E MTL framework with LID CSSR methods ($\lambda_2 = 0.2$). **'Man'**, **'En'** and **'ALL'** mean English, Mandarin and total TER of the test sets respectively.*

Systems	dev_{man} (%)			dev_{sge} (%)		
	Man	En	ALL	Man	En	ALL
E2E-DA-CHAR	21.8	39.2	26.5	28.1	44.2	38.4
+ LID_{shared}	21.8	38.7	26.3	27.7	43.6	37.9
+ LID_{indep}	**21.0**	38.0	**25.6**	**27.3**	42.4	37.0
E2E-DA-BPE3k	22.3	37.2	26.4	28.1	40.5	36.1
+ LID_{shared}	21.9	**37.0**	26.0	27.8	40.4	35.9
+ LID_{indep}	21.8	37.3	26.0	27.7	**40.3**	**35.8**

5.3. Effect of the NLM vocabulary expansion

Table 6 reports the TER results of the N-best (N=30) rescoring with the NLM vocabulary expansion method, where the NLM is the RNN-LM in practice. We see from Table 6 that the proposed NLM vocabulary expansion method achieved consistent improved TER results over the best results shown in the last row of Table 5. Finally, experiment results showed that applying the proposed approaches significantly reduces the TER of two dev sets from 34.5% and 46.5% to 25.0% and 34.5% respectively, which is close to the results of strong LF-MMI TDNN system.

Table 6: *TER of the NLM vocabulary expansion.*

Method	dev_{man} (%)	dev_{sge} (%)
NoLM	26.0	35.8
5-gram KN	25.9	35.6
RNN-LM	25.1	34.6
Proposed NLM	**25.0**	**34.5**

6. Conclusion and future work

In this paper we proposed several approaches to improve E2E based Mandarin-English code-switching speech recognition. These approaches include data augmentation, byte-pair encoding subword units for English language, language identification based multitask learning, as well as the vocabulary expansion for neural language models to rescore the N-best results.

In the future, we plan to study leveraging external monolingual data to improve its performance. we also plan to incorporating language model into the existing model to improve its performance.

7. Acknowledgment

This work is supported by the project of Alibaba-NTU Singapore Joint Research Institute.

8. References

[1] E. Yilmaz, H. van den Heuvel, and D. A. van Leeuwen, "Acoustic and textual data augmentation for improved ASR of code-switching speech," in *Proc. of INTERSPEECH*, 2018.

[2] P. Guo, H. Xu, L. Xie, and E. S. Chng, "Study of semi-supervised approaches to improving english-mandarin code-switching speech recognition," in *Proc. of INTERSPEECH*, 2018.

[3] A. Zeyer, K. Irie, R. Schlüter, and H. Ney, "Improved training of end-to-end attention models for speech recognition," *arXiv preprint arXiv:1805.03294*, 2018.

[4] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *arXiv preprint arXiv:1508.07909*, 2015.

[5] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," *arXiv preprint arXiv:1804.10959*, 2018.

[6] S. J. Pan, Q. Yang *et al.*, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[7] Y. Khassanov and E. S. Chng, "Unsupervised and efficient vocabulary expansion for recurrent neural network language models in asr," *Proc. of INTERSPEECH*, 2018.

[8] E. van der Westhuizen and T. Niesler, "Synthesising isizulu-english code-switch bigrams using word embeddings," in *Proc. of INTERSPEECH*, 2017.

[9] N. T. Vu, D.-C. Lyu, J. Weiner *et al.*, "A first speech recognition system for Mandarin-English code-switching conversational speech," in *Proc. of ICASSP*, 2012.

[10] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. of ICML*, 2006.

[11] W. Chan, N. Jaitly *et al.*, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. of ICASSP*, 2016.

[12] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. of ICML*, 2014.

[13] A. Y. Hannun, A. L. Maas, D. Jurafsky, and A. Y. Ng, "First-pass large vocabulary continuous speech recognition using bi-directional recurrent dnns," *arXiv preprint arXiv:1408.2873*, 2014.

[14] A. Hannun, C. Case, J. Casper *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[15] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014.

[16] C.-C. Chiu, T. N. Sainath, Y. Wu *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," *arXiv preprint arXiv:1712.01769*, 2017.

[17] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. 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.

[18] S. Watanabe, T. Hori, and J. R. Hershey, "Language independent end-to-end architecture for joint language identification and speech recognition," in *Proc. of ASRU*, 2017.

[19] S. Toshniwal, T. Sainath, R. Y. Weiss *et al.*, "Multilingual speech recognition with a single end-to-end model," in *Proc. of ICASSP*, 2018.

[20] P. Ramachandran, P. J. Liu, and Q. V. Le, "Unsupervised pretraining for sequence to sequence learning," *arXiv preprint arXiv:1611.02683*, 2016.

[21] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. of INTERSPEECH*, 2015.

[22] T. Ko, V. Peddinti, D. Povey *et al.*, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. of ICASSP*, 2017.

[23] D.-C. Lyu, T. P. Tan, E. Chng, and H. Li, "Seame: a mandarin-english code-switching speech corpus in south-east asia." in *Proc. of INTERSPEECH*, 2010.

[24] D. Povey, A. Ghoshal, G. Boulianne *et al.*, "The kaldi speech recognition toolkit," in *Proc. of ASRU*, 2011.

[25] D. Povey, V. Peddinti *et al.*, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. of INTERSPEECH*, 2016.

[26] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[27] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems 28*, 2015, pp. 577–585.