INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

End-to-End Spelling Correction Conditioned on Acoustic Feature for Code-switching Speech Recognition

Shuai Zhang[1,2], *Jiangyan Yi*[2], *Zhengkun Tian*[1,2], *Ye Bai*[1,2], *Jianhua Tao*[1,2,3], *Xuefei Liu*[2], *Zhengqi Wen*[2]

[1]School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[2]NLPR, Institute of Automation, Chinese Academy of Sciences, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, China

{shuai.zhang, jiangyan.yi, zhengkun.tian, ye.bai, jhtao, xuefei.liu, zqwen}@nlpr.ia.ac.cn

Abstract

In this work, we propose a new end-to-end (E2E) spelling correction method for post-processing of code-switching automatic speech recognition (ASR). Existing E2E spelling correction models take the hypotheses of ASR as inputs and annotated text as the targets. Due to the powerful modeling capabilities of the E2E model, the training of the correction system is extremely prone to over-fitting. It usually requires sufficient data diversity for reliable training. Therefore, it is difficult to apply the E2E correction models to the code-switching ASR task because of the data shortage. In this paper, we introduce the acoustic features into the spelling correction model. Our method can alleviate the problem of over-fitting and has better performance. Meanwhile, because the acoustic features are encode-free, our proposed model can be applied to the ASR model without significantly increasing the computational cost. The experimental results on ASRU 2019 Mandarin-English Code-switching Challenge data set show that the proposed method achieves 11.14% relative error rate reduction compared with baseline.

Index Terms: code-switching, speech recognition, end-to-end, spelling correction

1. Introduction

Code-switching is a common language phenomenon in which a sentence contains multiple languages [1]. Tt is difficult to obtain a lot of code-switching training data for automatic speech recognition (ASR). Data scarcity brings many challenges to the ASR system [2]. For the code-switching ASR task, the end-to-end (E2E) models develop rapidly because of their simplicity compared with the pipeline framework [3, 4, 5, 6, 7]. Although some progress has been made in the E2E code-switching ASR, training data shortage limits the further improvement of the models' performance.

E2E method integrates acoustic, pronunciation, and language models into a whole and is optimized jointly [8, 9, 10, 11, 12, 13, 14]. However, the E2E model does not perform well when the samples contain rare words which appear infrequently in the training set. It is because the E2E models only use speech-text paired data for training. A large amount of text data cannot be used. This problem is more serious in code-switching ASR task than monolingual.

In order to further improve the ASR performance, various post-processing techniques are used [15, 16, 17, 18, 19]. An effective method is to use a language model to re-score the recognized n-best hypotheses [9]. The external language model is usually trained independently on a large number of text corpora. The additional language model can effectively compensate for the text sparsity of the E2E model. Generally, the statistical language models and neural network language models re-scoring can effectively improve the accuracy of ASR. This technique is simple and effective, but it has some disadvantages. First, these left-to-right language models have an error accumulation problem, that is, if a recognition error occurs at the beginning of speech decoding, subsequent scoring will be misleading [19]. Second, The language models are not integrated into the E2E model with the objective of correcting errors made by the E2E model [18]. To alleviate the problems, some conditional language models are proposed to correct the errors generated by the recognition system [18, 19]. These spelling correction models take the hypotheses of ASR as input and annotated text as the target, similar to the process of neural machine translation (NMT) [20, 21]. By explicitly modeling the error patterns of the recognition model, the error correction model can improve the performance of the ASR system.

[18] proposes a spelling correction model based on RNN with attention for Listen, Attend and Spell (LAS) model [9]. It adopts a text-to-speech model to generate enough text-text paired training data. Similar to [18], [19] is based on the transformer model and uses data perturbation technology [22] to generate a large amount of training data. To alleviate the over-fitting problem, the pre-trained model BERT [23] is used to initialize the model parameters. The speech data they used is about 1000 hours [24], which is still difficult to obtain for code-switching. To address the problem, we introduce the acoustic information into the spelling correction model. Acoustic information and text information are fused through the attention mechanism. This method can provide more information for the decoding process. Our method can alleviate the problem of over-fitting and has better performance. Therefore, the spelling correction model can be applied to the code-switching ASR task more effectively. Meanwhile, because the acoustic features are encode-free, our proposed model can be applied to the ASR model without significantly increasing the computational cost. When compared to language model re-scoring methods and text-only spelling correction model using ASRU 2019 Mandarin-English code-switching Challenge data set [25], our method is more effective in improving the accuracy of the ASR system.

The rest of the paper is organized as follows. Section 2 describes our methods in detail, including the ASR model, language model, and error correction model. Section 3 introduces our experiment setups, including data augmentation, evaluation metrics, and models' parameters. In Section 4, we analyze the experimental results quantitatively and qualitatively in detail. Finally, we conclude our work in Section 5.

http://dx.doi.org/10.21437/Interspeech.2021-1242

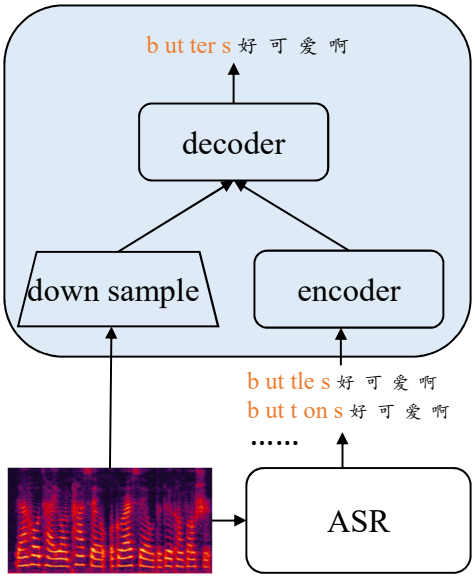

Figure 1: *The architecture of E2E spelling correction model conditioned on acoustic.*

2. Methods

2.1. Baseline speech recognition model

The baseline ASR model used for experiments in this work is speech-transformer [10], which is an attention based encoder-decoder model. We first briefly introduce the various modules of the model and their functions. For the encoder, a down-sampling module is usually used to reduce the number of speech features frames. Then a stack of multi-head self-attention based encoder-blocks is used to get the final encoded representations. It receives a sequence of acoustic features $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$ and transform them into intermediate representations $\mathbf{e} = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_{n'}]$, where n and n' are the initial and down-sampled frame numbers respectively. Then the encoder output is passed to an attention mechanism, which gets the soft alignments between the state of the decoder and input features frames. And a context vector $\mathbf{c}$ is obtained by performing a weighted average operation according to the attention weights. Finally, the decoder fusions the context vector $\mathbf{c}$ and the previous tokens to predict the current output. The decoder consists of a stack of self-attention and cross-attention based modules. The model trained by minimizing the cross-entropy loss on the training data.

2.2. External language model

A language model estimates the joint probability of a sentence $(y_1, y_2, ..., y_t)$ using a chain rule:

$$P(y_1, y_2, ..., y_t) = \prod_{t=1}^{T} P(y_t|y_1, ..., y_{t-1}) \quad (1)$$

Some neural network language models based on RNN or LSTM usually model this probability. To simplify modeling, it is assumed that the number of preceding words is limited to N:

$$P(y_t|y_1, ..., y_{t-1}) \approx P(y_t|y_{t-N}, ..., y_{t-1}) \quad (2)$$

This approximation method refers to the N-gram language model. These language models trained with large amounts of text can effectively improve the performance of the ASR model. There are several main mechanisms to use external language model for ASR model [15, 16, 17, 26]. [9] shows significant improvement by simply using a language model to re-score the n-best hypotheses decoded by the ASR model. Several language model fusion mechanisms that operate step by step in the decoding stage are also effective for ASR task [17, 26]. In this work, we focus on the n-best re-scoring method. It can be easily compared with our method to evaluate the effectiveness of the proposed method.

2.3. Spelling correction model with acoustic features

The above-mentioned external language models are all independently trained using text corpus, and what they learn is general language information. However, they cannot take into account the error patterns of a specific ASR system. Therefore, some spelling correction models take the hypotheses of ASR as input and annotated text as the target to explicitly model the error distribution. From a certain point of view, the spelling correction model is a kind of conditional language model. For the text-only correction model, it can be formalized as:

$$P(y_1, y_2, ..., y_t) = \prod_{t=1}^{T} P(y_t|y_1, ..., y_{t-1}, C_{text}) \quad (3)$$

where the C_{text} refers to the error outputs of the ASR model. However, the input text and output text of the model are mostly similar due to the high accuracy of the recognition model. Meanwhile, deep neural networks have powerful modeling capabilities, this task is prone to over-fitting. In this paper, we add acoustic features as additional prior information in the spelling correction model.

$$P(y_1, y_2, ..., y_t) = \prod_{t=1}^{T} P(y_t|y_1, ..., y_{t-1}, C_{text}, C_{acoustic}) \quad (4)$$

Intuitively, the model fusions the error distribution information C_{text} and acoustic information $C_{acoustic}$ at the same time.

Our error correction method is based on the transformer model, an efficient attention based encoder-decoder model. As shown in Figure 1, the ASR system first output a list of hypotheses, then the encoder receives these hypotheses and transforms them into continuous representations. In addition to text input, the acoustic features also input into the decoder after down-sampling. The down-sampling technique splices several adjacent features frames without generating trainable parameters. Then the decoder performs an attention query operation on the two input sequences to generate context vectors respectively. The two kinds of context vectors are concatenated to fusion the text and acoustic information. As the acoustic features are encode-free, our proposed model can be applied to a variety of different ASR models without significantly increasing the computational cost.

3. Experiments

3.1. Datasets

We conduct our experiments on ASRU 2019 Mandarin-English code-switching Challenge dataset [25], which consists of about

500 hours of Mandarin data and 200 hours of code-switching training data. In this paper, we only use the code-switching data to generate the error correction training data. The development set and test set each has about 20 hours of code-switching data. For the code-switching data, Mandarin is the host language and English is the guest language. All speech data is collected in quiet rooms by Android or IOS phones with 16kHz sampling rate. The transcripts of data cover many common fields including entertainment, travel, daily life, and social interaction.

3.2. Data augmentation

To generate the training data for the transformer spelling correction model, we split all training data into 10 folds and trained 10 ASR baseline model in a cross-validation manner: 9 folds data are used to train a ASR model and the remaining 1 fold data is used to generate ASR hypotheses using beam search. The beam size is set to 10. Then we get 10 times the data of the original text. Due to the problem of over-fitting, we cannot train an effective error correction model using these data. Therefore, we add perturbation techniques such as SpecAugment [27] and dropout [28] in the model inference process, and repeat this process many times using different random seeds and parameters. In order to eliminate interference, we removed all samples whose decoding accuracy exceeded 90%. Finally, we get about 2M of training data after removing duplication.

3.3. Evaluation metric

Because our goal is to improve the performance of the code-switching ASR model, we adopt the error rate metric to evaluate the spelling correction model. In this paper, mix error rate (MER) is used to evaluate the experimental results. The MER is defined as the word error rate (WER) for English and character error rate (CER) for Mandarin. This metric can balance the Mandarin and English error rates better compared to the WER or CER. The metric is widely adopted to evaluate the performance of Mandarin-English code-switching ASR system.

3.4. Experimental setups

In this subsection, we introduce some important parameter setups of our methods, including the ASR model, language model, and spelling correction model with the acoustic features.

ASR: The input features of the acoustic encoder network are 40-dimensional Mel filter-bank with 25ms windowing and 10ms frame shift using Kaldi toolkit [29]. Two $3 * 3$ 2D CNN down-sampling layers with stride 2 for the acoustic features are used. The attention dimensions of the encoder and decoder are both 256, and the number of heads is 4. The dimension of position-wise feed-forward networks is 1024. And the number of acoustic encoder blocks and decoder blocks are 12 and 6 respectively.

Language model: In order to evaluate our method more comprehensively, we use two kinds of language models to re-score the ASR results. The first is the N-gram model. This kind of language model is sufficiently effective and efficient. Specifically, the 6-gram KenLM [31] is used. And the second is a traditional left-to-right neural network language model. Its structure includes two unidirectional LSTM layers with a dimension of 1024. We use the code-switching transcript to train the external language model. The modeling units is the same as the ASR model. This re-score process can be formally expressed as follows:

$$y^* = argmax_y logP(y|x) + \lambda logP_{LM}(y) \quad (5)$$

where $P(y|x)$ and $P_{LM}(y)$ refer to scores from ASR model and language model. And the λ is the hyper-parameter that determines the weight of the language model. We set it to 0.5 in the experiments.

spelling correction: The input acoustic features of the correction model are the same as the ASR model. The down-sampling module splices 10 adjacent features frames without generating trainable parameters. The attention dimensions of the encoder and decoder are both 256, and the number of heads is 4. The dimension of position-wise feed-forward networks is 512. And the number of acoustic encoder blocks and decoder blocks are 6 and 6 respectively. The decoder uses a dimension conversion layer to unify the dimensions of text representations and acoustic features. The uniform label smoothing technique is used and the parameter is set to 0.1 [32]. Meanwhile, we set residual dropout as 0.1, where the residual dropout is applied to each sub-block before adding the residual information. We use Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.98$ [33]. The learning rate is set by a warm-up strategy. After training, we average the last 5 checkpoints as the final model. And we perform decoding using beam search with a beam size of 10. In order to verify the necessary of acoustic features in the error correction model, we also train a text-only error correction model. Its parameters are the same as the above setups.

For the output target, we adopt Chinese characters and English word pieces as the modeling units. We select the characters which appear more than five times in the training set as the Chinese modeling units. And the number of English word pieces is 1k [30]. The word pieces cannot only alleviate the out-of-vocabulary (OOV) problem with limited English training data but also balance the granularity of Chinese and English modeling units.

4. Results

Table 1: *The MER/CER/WER (%) of several post-processing methods. CH refers to the CER of the Chinese part in code-switching. EN refers to the WER of the English part in code-switching. All refers to the MER of the code-switching.*

Model	Dev			Test		
	All	CH	EN	All	CH	EN
ASR	12.67	10.33	31.52	11.94	9.71	30.24
+ 6-gram LM	11.97	9.89	28.70	11.37	9.35	27.93
+ LSTM LM	11.71	9.61	28.66	11.21	9.17	27.97
Text-only	12.61	9.57	28.10	11.24	9.26	27.51
Our	11.22	9.18	27.66	**10.61**	8.64	26.73

4.1. Results of the spelling correction

First, we quantitatively compare our method's results with several other methods. AS shown in Table 1, for the ASR baseline model, the error rate of English is much greater than that of Chinese. It is consistent with the results in other papers. This may be because Mandarin is the host language and English is the guest language in the data set. Then we use the language model to re-score the recognition results, and we find that the MER is reduced. Specifically, the neural network language model achieves better performance compared to the N-gram language model. We find that the language model re-scoring technology improves the accuracy of English more significantly than

Table 2: *Samples produced by different methods. We select two samples which beam search decoding and re-scoring with external LM fail to recognized correctly. Some rare combinations of Chinese and English words and similar pronunciation caused errors, our model can correct them completely or partially.*

Model	Example 1	Example 2
Ground truth	net core 微服务脚手架然后开源	news 兔和 shower 狗相爱四十九天
ASR	night crawl 微服务脚手架然后开员	news two 和沙窝 go 相爱四十九天
+ 6-gram LM	night crawl 微服务脚手架然后开源	news to 和杀我 go 相爱四十九天
+ LSTM	night crawl 微服务脚手架然后开源	news to 和杀我 go 相爱四十九天
Text-only	light core 微服务脚手架然后开源	news to 和 shower go 相爱四十九天
Ours	net core 微服务脚手架然后开源	news to 和 shower 狗相爱四十九天

Table 3: *Samples produced by spelling correction models which trained with the text decoded by the monolingual ASR model. English words are recognized as combinations of Chinese words with similar pronunciations.*

Model	Example 1
Ground truth	means 抠开看发现没熟
ASR	密斯抠开看发现没熟
Text-only	means coke 看发现没熟
Ours	means 抠开看发现没熟

that of Chinese. The possible reason is that the language model based on sub-words can correct many spelling errors of English words. And the performance of the text-only spelling correction model is similar to that of re-scoring operations. The limited diversity of training text limits the performance of the error correction model. The model cannot learn more error distribution patterns. Finally, our proposed spelling correction model with acoustic information achieves the best performance. These results fully demonstrate that acoustic information can effectively assist in the error correction process. Overall, the proposed model provides up to 11.14% relative reduction in MER compared with the baseline. And English WER and Chinese CER are improved 11.33% and 11.61% respectively. Meanwhile, the proposed model provides up to 5.60% relative reduction in MER compared with the text-only correction model.

4.2. Error analysis

In addition to quantitative analysis, we also conduct qualitative analysis for some examples. Some interesting phenomena can be found in this way. For example 1 in Table 2, the ASR model has recognition errors in both the Chinese and English parts. These wrong words are the same or similar in pronunciation as the target words. After the re-scoring, the common words in the Chinese part can be corrected. However, the uncommon word combinations in the English part have not been corrected. This left-to-right language model cannot make good use of global contextual information. The text-only spelling correction model can correct part errors according to the context. However, with the aid of acoustic information, our method can get exactly the right target. For example 2, English words are mistakenly recognized as Chinese words with similar pronunciations. Since the combination of Chinese and English words in the text is relatively rare, the language model is helpless. It can even generate new errors. By considering all the context information, the error correction model has better performance. The introduction of acoustic information further enhances the ability of the error correction model.

4.3. Correction on monolingual speech recognition

Further, we consider the case of using a monolingual recognition system to recognize code-switching speech. The recognition error generated in this situation should also be corrected intuitively. So we use the Chinese ASR model which trained with 500 hours of monolingual speech data. We use the same data augmentation methods as previously described to generate spelling correction data. Then the correction models with or without acoustic information are trained. Unfortunately, we did not get meaningful quantitative results. There should be several reasons for this. The first is the lack of training data, and the second is that the pronunciation unit of Chinese characters cannot model the pronunciation of English words well. Fortunately, we get some qualitative results. Table 3 shows a specific error correction sample. It is obvious that English words are recognized as a combination of Chinese characters with similar pronunciation. Text-only correction model corrects some errors, but at the same time produces similar error pattern. This illustrates that it is difficult for the text-only correction model to decide whether to correct errors or not. This problem can be alleviated to a certain extent with the help of acoustic information. After obtaining a large amount of code-switching data, the correction model for the monolingual system may be a feasible method.

5. Conclusions and future work

In this work, we propose a new E2E spelling correction method for post-processing of code-switching ASR. Compared to the text-only correction models, we introduce the acoustic information into the spelling correction model. Our method can alleviate the problem of over-fitting and has better performance. Meanwhile, because the acoustic features are encode-free, our proposed model can be applied to the ASR model without significantly increasing the computational cost. When compared to language model re-scoring methods and text-only spelling correction models, our method is more effective in improving the accuracy of the ASR system.

6. Acknowledgment

This work is supported by the National Key Research & Development Plan of China (No.2017YFC0820602), the National Natural Science Foundation of China (NSFC) (No.61831022, No.61901473, No.61771472, No.61773379), and Inria-CAS Joint Research Project (No.173211KYSB20190049). And this research is (partially) funded by Huawei Noah's Ark Lab.

7. References

[1] P. Muysken, P. C. Muysken *et al.*, *Bilingual speech: A typology of code-mixing*. Cambridge University Press, 2000.

[2] Ö. Çetinoğlu, S. Schulz, and N. T. Vu, "Challenges of computational processing of code-switching," *arXiv preprint arXiv:1610.02213*, 2016.

[3] Z. Zeng, Y. Khassanov, V. T. Pham, H. Xu, E. S. Chng, and H. Li, "On the end-to-end solution to mandarin-english code-switching speech recognition," *arXiv preprint arXiv:1811.00241*, 2018.

[4] C. Shan, C. Weng, G. Wang, D. Su, and L. Xie, "Investigating end-to-end speech recognition for mandarin-english code-switching," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[5] K. Li, J. Li, G. Ye, R. Zhao, and Y. Gong, "Towards code-switching asr for end-to-end ctc models," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6076–6080.

[6] S. Zhang, J. Yi, Z. Tian, J. Tao, and Y. Bai, "Rnn-transducer with language bias for end-to-end mandarin-english code-switching speech recognition," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2021, pp. 1–5.

[7] S. Zhang, J. Yi, Z. Tian, Y. Bai, J. Tao *et al.*, "Decoupling pronunciation and language for end-to-end code-switching automatic speech recognition," *arXiv preprint arXiv:2010.14798*, 2020.

[8] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.

[9] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[10] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5884–5888.

[11] Z. Tian, J. Yi, J. Tao, Y. Bai, and Z. Wen, "Self-attention transducers for end-to-end speech recognition," *Proc. Interspeech 2019*, pp. 4395–4399, 2019.

[12] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4945–4949.

[13] Y. Zhang, W. Chan, and N. Jaitly, "Very deep convolutional networks for end-to-end speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4845–4849.

[14] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[15] J. Chorowski and N. Jaitly, "Towards better decoding and language model integration in sequence to sequence models," *Proc. Interspeech 2017*, pp. 523–527, 2017.

[16] A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, and R. Prabhavalkar, "An analysis of incorporating an external language model into a sequence-to-sequence model," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5828.

[17] S. Toshniwal, A. Kannan, C.-C. Chiu, Y. Wu, T. N. Sainath, and K. Livescu, "A comparison of techniques for language model integration in encoder-decoder speech recognition," in *2018 IEEE spoken language technology workshop (SLT)*. IEEE, 2018, pp. 369–375.

[18] J. Guo, T. N. Sainath, and R. J. Weiss, "A spelling correction model for end-to-end speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5651–5655.

[19] O. Hrinchuk, M. Popova, and B. Ginsburg, "Correction of automatic speech recognition with transformer sequence-to-sequence model," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7074–7078.

[20] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *arXiv preprint arXiv:1409.3215*, 2014.

[21] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *EMNLP*, 2014.

[22] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[25] X. Shi, Q. Feng, and L. Xie, "The ASRU 2019 mandarin-english code-switching speech recognition challenge: Open datasets, tracks, methods and results," *CoRR*, vol. abs/2007.05916, 2020. [Online]. Available: https://arxiv.org/abs/2007.05916

[26] A. Sriram, H. Jun, S. Satheesh, and A. Coates, "Cold fusion: Training seq2seq models together with language models," *Proc. Interspeech 2018*, pp. 387–391, 2018.

[27] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[30] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *EMNLP (Demonstration)*, 2018.

[31] K. Heafield, "Kenlm: Faster and smaller language model queries," in *Proceedings of the sixth workshop on statistical machine translation*, 2011, pp. 187–197.

[32] R. Müller, S. Kornblith, and G. Hinton, "When does label smoothing help?" *arXiv preprint arXiv:1906.02629*, 2019.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.