INVESTIGATING END-TO-END SPEECH RECOGNITION FOR MANDARIN-ENGLISH CODE-SWITCHING

Changhao Shan[1,2*], Chao Weng[4], Guangsen Wang[3], Dan Su[3], Min Luo[2], Dong Yu[4], Lei Xie[1†]

[1]School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, China
[2]Tencent AI Platform Department, Shenzhen, China
[3]Tencent AI Lab, Shenzhen, China
[4]Tencent AI Lab, Bellevue, USA

{chshan, lxie}@nwpu-aslp.org, {cweng, vincegswang, dansu, selwynluo, dyu}@tencent.com

ABSTRACT

Code-switching is a common phenomenon in many multilingual communities and presents a challenge to automatic speech recognition (ASR). In this paper, three approaches are investigated to improve end-to-end speech recognition on Mandarin-English code-switching task. First, multi-task learning (MTL) is introduced which enables the language identity information to facilitate Mandarin-English code-switching ASR. Second, we explore wordpieces, as opposed to graphemes, as English modeling units to reduce the modeling unit gap between Mandarin and English. Third, we employ transfer learning to utilize larger amount of monolingual Mandarin and English data to compensate the data sparsity issue of a code-switching task. Significant improvements are observed from all three approaches. With all three approaches combined, the final system achieves a character error rate (CER) of 6.49% on a real Mandarin-English code-switching task.

Index Terms— automatic speech recognition, end-to-end speech recognition, attention-based model, code-switching

1. INTRODUCTION

Code-switching speech is defined as speech which contains more than one language. There are two different forms of code-switching: inter-sentential and intra-sentential. The phenomenon of code-switching is quite common around the world, for example between Mandarin and English [1, 2, 3, 4, 5, 6, 7], French and Algerian [8], Spanish and English [9]. Particularly, Mandarin-English code-switching is extremely frequent and popular in East Asia. In this study, we focus on intra-sentential code-switching that is both common and challenging.

The research work is supported by the National Key Research and Development Program of China (No.2017YFB1002102) and Tencent AI Lab Rhino-Bird Joint Research Program (No.JR201853).

*Work performed during an internship at Tencent.

†Corresponding author.

Building a code-switching ASR system has several challenges. The major one is the lack of publicly-available code-switching data as there exists only a few small corpus [8, 10, 11]. Meanwhile, the code-switching data have a highly unbalanced language distribution, especially for intra-sentential code-switching task. To address the unbalanced distribution issue, units merging [1, 2, 4] was employed to compensate the training data for the weak language. For language modeling, text augmentation approach was used to obtain the extra text data, such as the use of statistical machine translation [1] and word embedding [12]. Furthermore, other language specific information such as part-of-speech has shown to improve the performance of language model for code-switching as well [6, 13]. Another challenge for the conventional ASR based code-switching system is the need of hand-crafted language-dependent components, such as a bi-lingual phone set and a pronunciation lexicon. However, the modeling unit gap between languages imposes another challenge for the code-switching task due to the significant language discrepancy.

Recently, attention-based end-to-end model has proven to be effective in various speech tasks, such as ASR [14, 15, 16, 17, 18], keyword spotting [19] and speaker verification [20]. For code-switching task, Seki et al. [21] introduced a hybrid attention/CTC model that can recognize the mixed-language speech and the main advantage is that the need of handmade language-dependent resources was avoided. A simulated dataset that was generated by concatenating existing monolingual dataset was used for inter-sentential code-switching task. In addition, the system used the language identify explicitly and the joint language identification (ID) [21, 17] was employed to predict the code-switching between languages. Finally, with a new training procedure, this work achieved a competitive performance.

In this paper, to avoid the need of handmade resources, we adopt attention-based end-to-end model for the Mandarin-English code-switching speech recognition task. Specifically, we propose three improvements to the attention-based mod-

el for code-switching speech recognition. Firstly, multi-task learning [18] is introduced to enable the language identity information to facilitate Mandarin-English code-switching ASR. Specifically, three different structures to apply language ID loss are explored. Secondly, we explore wordpieces [22], as opposed to graphemes, as English modeling units to reduce the modeling unit gap between Mandarin and English. In addition, wordpieces can capture more context information and reduce the decoding time. Finally, we employ transfer learning [23] to further improve the performance by using the larger size of Mandarin and English monolingual data. Experiments on a 1K-hrs Mandarin-English code-switching dataset demonstrate that significant improvements are observed from all three proposed approaches and the final system achieves a CER as low as 6.49%.

The rest of this paper is organized as follows. Section 2 briefly describes our baseline architecture of attention-based end-to-end ASR. In Section 3, we detail the proposed improvements for our baseline model. We present our experiments and results in Section 4. Finally, we conclude our work in Section 5.

2. ATTENTION-BASED MODEL

2.1. Listen, Attend and Spell

Listen, Attend and Spell (LAS) [14] is an attention-based encoder-decoder network for end-to-end speech recognition. As depicted in Fig. 1-a, an LAS model consists of three components: an encoder network (the Listen module), a decoder network (the Spell module) and an attention model (The Attend module). The encoder network transforms the input feature $\mathbf{x} = (x_1, ..., x_T)$ into a high level representation $\mathbf{h^{enc}} = (h_1^{enc}, ..., h_T^{enc})$:

$$\boldsymbol{h^{enc}} = Encoder(\boldsymbol{x}). \tag{1}$$

The decoder network outputs the hidden state h_t^{dec} using the previous target label y_{t-1} and context information c_{t-1}:

$$h_t^{dec} = Decoder(y_{t-1}, c_{t-1}). \tag{2}$$

The attention model weights the feature representation and form a fixed-length vector c_t:

$$c_t = Attend(\boldsymbol{h^{enc}}, h_t^{dec}). \tag{3}$$

Then the attentional hidden state h_t^{att} is obtained using information from hidden state h_t^{dec} and context information c_t:

$$h_t^{att} = tanh(\boldsymbol{W_h}[c_t; h_t^{dec}]). \tag{4}$$

Finally, the output distribution is produced by a projection and the softmax layer:

$$y_t = softmax(\boldsymbol{W_o} h_t^{att}) \tag{5}$$

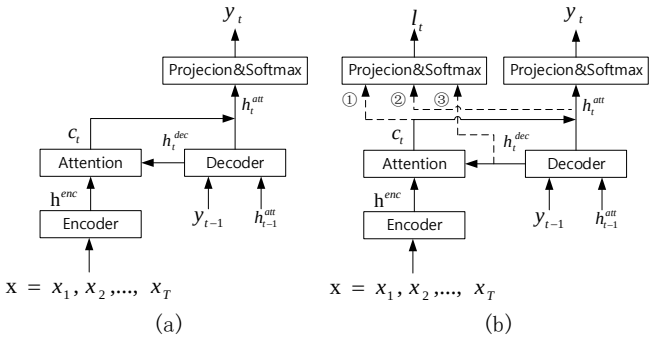

Fig. 1. Basic LAS model (a) and multi-task learning with language ID (b).

In this work, we adopt content-based attention [24], which is described by the following equations:

$$c_t = \sum_{i=1}^{T} \alpha_i h_i^{enc}, \tag{6}$$

$$\alpha_i = \frac{exp(e_i)}{\sum_{j=1}^{T} exp(e_j)}, \tag{7}$$

$$e_i = v_a^T tanh(\boldsymbol{W_a}[h_t^{dec}, h_i^{enc}]). \tag{8}$$

2.2. Input-feeding

In Eq. (2), the inputs of decoder are y_{t-1} and c_{t-1}. One shortcoming of this is the lack of information about previous alignment decisions. To address this, we employ the input-feeding architecture [16, 25] to feed the previous attentional hidden state h_{t-1}^{att} to the decoder. The hidden state h_t^{dec} is obtained as:

$$h_t^{dec} = Decoder(y_{t-1}, h_{t-1}^{att}). \tag{9}$$

2.3. Softmax smoothing

An attention-based model tends to give over-confident predictions. Softmax smoothing is applied to smooth the distribution of label prediction during decoding [15, 26]. The prediction probability is smoothed by a temperature hyperparameter $\tau = 2$:

$$y_t = softmax(\boldsymbol{W_o} h_t^{dec}/\tau). \tag{10}$$

3. METHODS

3.1. Language identity

In [1], the language ID classification at frame level achieved limited improvement for speech recognition performance. As shown in [21], language ID was useful for code-switching task and the joint language ID was employed. Specifically, the joint language ID was implemented by introducing language index in the beginning of the output text. This means that the

joint language ID only needs to predict the code-switching between languages. Different from the joint language ID, at each time step, we used multi-task learning to predict both the modeling unit and the language ID l_t. The language ID is treated as the secondary task and the language ID was predicted at character level. In addition, we explored three locations where the language ID loss can be applied when doing multi-task learning as shown in Fig. 1-b:

①:The output of attention model which contains the acoustic information. (c_t)

②:The attentional hidden state which contains the information of label prediction. (h^{att})

③:The output of decoder network which contains the language information. (h^{dec})

3.2. Wordpieces

For Mandarin-English code-switching task, a natural way to construct modeling units is to combine the Chinese characters with English letters as an augmented character set [21]. However, there is a significant discrepancy between Chinese characters and English letters. The acoustic counterpart of a Chinese character is longer than an English letter. We propose to adopt wordpieces as English modeling units. The motivation is two-fold: 1) increasing the duration of English modeling units will reduce the modeling unit gap between Mandarin and English. 2) for attention-based English ASR, compared to whole word unit, the wordpiece unit is proven to both improve the performance and resolve the OOV issue [22].

3.3. Transfer learning

To contrast to the limited Mandarin-English code-switching data, there are plenty of monolingual training data for both Mandarin and English. It is well known that the attention based ASR system is usually data hungry. It is interesting to investigate how to utilize the monolingual data to improve the system further. It's high likely that the model trained on monolingual data can be improved on individual languages but might have troubles to predict the switch between Mandarin and English. Following the transfer learning framework [23], we first train the LAS model using both code-switching and monolingual data for a better initialization, especially for the encoder network. Subsequently, we retrain the LAS model using the Mandarin-English code-switching data. This helps the LAS model to learn more information on the language switch.

4. EXPERIMENTS

4.1. Dataset

Three dataset were used for the study: 1) Mandarin-English code-switching data set (Mn-En), 2) monolingual Mandarin data set, 3) monolingual English data set. The Mn-En data set contains about 1K hours speech data, which has ∼840K utterances. In total, the dataset contains 6,079 unique Mandarin characters and 10921 unique English words. Besides, most of the utterances has one English word and some code-switching utterance is shown in Table 1. The ratio of Mandarin characters and English words are 88.9% and 11.1% respectively. The Mandarin monolingual data set has 4K hours speech data, which contains ∼4,235K utterances. The English monolingual data set consists of about 400 hours speech data, which has ∼487K utterances. We randomly selected 10,376 code-switching utterances (∼12.4 hours) as the test data. The speaker of test set has no overlap with the training and validation sets. Each audio frame was computed based on a 80-channel Mel-filterbank with 25ms windowing and 10ms frame shift and central frame was spliced with left 3 plus right 3 frames. Mean and variance normalization was conducted on the speaker level. The targets of our end-to-end system are a set of 6640 characters which contains English letters, English wordpieces, Mandarin characters, punctuations plus '<space>', '<SOS>' and '<EOS>'. Meanwhile, the language-independent output layer [27] was employed in our work.

Table 1. *Examples of Mandarin-English conde-switching utterance.*

那男孩因为hunger快死了.
(That boy is dying because of hunger.)
我现在想喝点milk.
(I want to drink some milk now.)
她离开了这个company了.
(She left the company.)

4.2. Experimental Setup

We constructed a common attention-based ASR system as described in Section 2. The encoder is a 6 layer BLSTM with 512 LSTM units per-direction (or 1024 in total). The Decoder is a 2 layer LSTM with 1024 LSTM units. We use ADAM [28] algorithm as the optimization method while we set β_1 = 0.9,β_2 = 0.999, $\epsilon = 10^{-8}$ as suggested in [28]. During training, we use 0.001 as initial learning rate and halve it if there is no improvement on the validation set. To reduce the length of the input sequences, we subsample the input by a factor of three similar to [16]. Meanwhile, to reduce overfitting, we use dropout [29] and set the value to 0.2 throughout all our experiments. Performances are measured by character error rate (CER) for Mandarin and word error rate (WER) for English. Due to the unbalanced language distribution, the overall performance can be roughly viewed as CER.

4.3. Results

From Table 2, we can clearly see that the baseline LAS architecture already achieves a good result. This means

Table 2. *Performance for Mn-En code-switching data set.*

Model	CER (%)
LAS	8.15
joint language ID [21]	8.97
wordpiece	**7.70**
MTL (c_t)	**7.88**
MTL (h^{att})	8.05
MTL (h^{dec})	8.87
MTL (c_t) + wordpiece	**7.60**
MTL (h^{att}) + wordpiece	7.67
MTL (h^{dec}) + wordpiece	7.93

Table 3. *Performance of language identification with Mn-En code-switching data set.*

Model	LER (%)
MTL (c_t)	**1.78**
MTL (h^{att})	1.82
MTL (h^{dec})	2.15
MTL (c_t) + wordpiece	**1.65**
MTL (h^{att}) + wordpiece	2.90
MTL (h^{dec}) + wordpiece	3.17

that by using a big dataset, we can learn a promising model to address the intra-sentential code-switching task. We also explored the performance of joint language ID [21]. However, due to the difference between inter-sentential and intra-sentential code-switching task, we were not able to obtained performance gain over our baseline. In inter-sentential code-switching task, predicting the code-switching between languages is easy and helpful. However, the intra-sentential code-switching is spontaneous and unpredictable and the code-switching between languages is difficult to predict. We analyzed that this leads to the worse performance of the joint language ID. Instead, we employed multi-task learning to utilize the language ID information in our work. At each time step, we predicted both the modeling unit and the language ID. In Table 2, we can see that both language ID with c_t and h^{att} are preferred over the baseline system. However, language ID with h^{dec} preforms worst than the baseline. This probably means that the encoder and attention modules capture most of the language ID information, which benefits the overall system performance. This also means that it is difficult to identify a language by relying only on the language model information in h^{dec}. From Table 3, we noticed that our models achieve very strong language ID recognition performance. The language ID with c_t presents the best performance and achieves a language ID error rate (LER) of 1.78%. We also observed that there is a strong positive correlation between the performance of speech recognition and language ID recognition.

We employed wordpieces as the English modeling units to reduce the difference between Mandarin and English. Moreover, the wordpieces contain more context information and reduce the length of English word. From Table 2, we can see that the wordpieces achieve a better performance than the usage of language ID information. Meanwhile, the language ID information can also achieve a marginal improvement, when used along with the wordpiece. In Table 3, similar to language ID, the combination of the wordpiece and language ID leads to significant performance. In addition, the strong positive correlation between speech recognition and language ID recognition can also be observed in Table 3.

Table 4. *Performance of adding Mandarin and English data.*

Model	Data	CER (%)
LAS	Mn-En + 4k hrs Mn + 400 hrs En	7.51
+ retrain	Mn-En	7.01
MTL (c_t)	Mn-En + 4k hrs Mn + 400 hrs En	7.09
+ retrain	Mn-En	**6.49**
wordpiece	Mn-En + 4k hrs Mn + 400 hrs En	7.73
+ retrain	Mn-En	6.87

Lastly, we explored the effect of supplementing the training data with monolingual Mandarin and English data. Experiments show that simply adding the monolingual data to the original code-switching data presents a significant advantage. From Table 4, the baseline architecture can achieve a CER of 7.51%. We believe that although the Mandarin and English speech data can better initialize the encoder network, the additive monolingual data increase the difficulty of predicting the switch between Mandarin and English. Therefore, we further employed transfer learning and retrained the model with Mn-En data which can further reduce the CER to 7.01% as shown in Table 4. Besides, both language ID and wordpiece can further improve the performance of speech recognition. Finally, our best model achieves a CER of 6.49%, a CER reduction of 13.5% compared to the baseline.

5. CONCLUSIONS

In this work, we investigated the attention-based end-to-end speech recognition for Mandarin-English code-switching. We achieved a promising result on a real Mandarin-English code-switching dataset. Using the wordpieces as the English modeling units, we reduced the gap of modeling units. We further utilized the language ID information to improve the recognition performance. In addition, transfer learning was employed to effectively use the monolingual Mandarin and English data and achieved a significant advantage. Finally, on a real Mandarin-English code-switching dataset, we achieved a CER of 6.49% for the attention-based end-to-end speech recognition.

6. REFERENCES

[1] 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 *ICASSP2012*.

[2] C.-F. Yeh and L.-S. Lee, "Transcribing code-switched bilingual lectures using deep neural networks with unit merging in acoustic modeling," in *ICASSP2014*.

[3] Y. Li, P. Fung, P. Xu, and Y. Liu, "Asymmetric acoustic modeling of mixed language speech," in *ICASSP2011*.

[4] C.-F. Yeh, C.-Y. Huang, and L.-S. Lee, "Bilingual acoustic model adaptation by unit merging on different levels and cross-level integration," in *Twelfth Annual Interspeech*, 2011.

[5] C.-F. Yeh, L.-C. Sun, C.-Y. Huang, and L.-S. Lee, "Bilingual acoustic modeling with state mapping and three-stage adaptation for transcribing unbalanced code-mixed lectures," in *ICASSP2011*.

[6] 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 *ICASSP2013*.

[7] P. Guo, H. Xu, L. Xie, and E. S. Chng, "Study of semi-supervised approaches to improving english-mandarin code-switching speech recognition," in *Interspeech2018*.

[8] D. Amazouz, M. Adda-Decker, and L. Lamel, "Addressing code-switching in french/algerian arabic speech," in *Interspeech2017*.

[9] G. Guzmán, J. Ricard, J. Serigos, B. E. Bullock, and A. J. Toribio, "Metrics for modeling code-switching across corpora," in *Interspeech2017*.

[10] D.-C. Lyu, T.-P. Tan, E.-S. Chng, and H. Li, "An analysis of a mandarin-english code-switching speech corpus: Seame," in *Interspeech2010*.

[11] J. Y. Chan, P. Ching, and T. Lee, "Development of a cantonese-english code-mixing speech corpus," in *Eurospeech2005*.

[12] E. van der Westhuizen and T. Niesler, "Synthesising isizulu-english code-switch bigrams using word embeddings," *Interspeech2017*.

[13] 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 ACL2013*, vol. 2.

[14] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell," in *ICASSP2016*.

[15] C. Shan, J. Zhang, Y. Wang, and L. Xie, "Attention-based end-to-end speech recognition on voice search," in *ICASSP2018*.

[16] C. Weng, J. Cui, G. Wang, J. Wang, C. Yu, D. Su, and D. Yu, "Improving attention based sequence-to-sequence models for end-to-end english conversational speech recognition," in *Interspeech2018*.

[17] S. Watanabe, T. Hori, and J. R. Hershey, "Language independent end-to-end architecture for joint language identification and speech recognition," in *ASRU2017*.

[18] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein, and K. Rao, "Multilingual speech recognition with a single end-to-end model," in *ICASSP2018*.

[19] C. Shan, J. Zhang, Y. Wang, and L. Xie, "Attention-based end-to-end models for small-footprint keyword spotting," in *Interspeech2018*.

[20] F. Chowdhury, Q. Wang, I. L. Moreno, and L. Wan, "Attention-based models for text-dependent speaker verification," *arXiv preprint arXiv:1710.10470*, 2017.

[21] H. Seki, S. Watanabe, T. Hori, J. Le Roux, and J. R. Hershey, "An end-to-end language-tracking speech recognizer for mixed-language speech," in *ICASSP2018*.

[22] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *ICASSP2018*.

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

[24] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[25] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *EMNLP2015*.

[26] J. Chorowski and N. Jaitly, "Towards better decoding and language model integration in sequence to sequence models," *arXiv preprint arXiv:1612.02695*, 2016.

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

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

[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *JMLR2014*, vol. 15, no. 1, pp. 1929–1958.