Towards Language-Universal Mandarin-English Speech Recognition

Shiliang Zhang[1], *Yuan Liu*[1], *Ming Lei*[1], *Bin Ma*[1], *Lei Xie*[2]

[1]Machine Intelligence Technology, Alibaba Group
[2]School of Computer Science, Northwestern Polytechnical University
{sly.zsl, hanyuan.ly, lm86501, b.ma}@alibaba-inc.com, lxie@nwpu.edu.cn

Abstract

Multilingual and code-switching speech recognition are two challenging tasks that are studied separately in many previous works. In this work, we jointly study multilingual and code-switching problems, and present a language-universal bilingual system for Mandarin-English speech recognition. Specifically, we propose a novel bilingual acoustic model, which consists of two monolingual system initialized subnets and a shared output layer corresponding to the *Character-Subword* acoustic modeling units. The bilingual acoustic model is trained using a large Mandarin-English corpus with CTC and sMBR criteria. We find that this model, which is not given any information about language identity, can achieve comparable performance in monolingual Mandarin and English test sets compared to the well-trained language-specific Mandarin and English ASR systems, respectively. More importantly, the proposed bilingual model can automatically learn the language switching. Experimental results on a Mandarin-English code-switching test set show that it can achieve 11.8% and 17.9% relative error reduction on Mandarin and English parts, respectively.

Index Terms: speech recognition, Mandarin-English, code-switching, bilingual, DFSMN-CTC-sMBR

1. Introduction

As voice-driven interface to smart devices become mainstream, increasing the language coverage of speech recognition systems is particularly important. There exists thousands of languages in human speech interaction, including various official-languages and different dialects. Usually, language-specific automatic speech recognition (ASR) system is built for each language, which requires a large number of wave-transcription pairs to train an acoustic model, huge amounts of text data to train a language model and tremendous linguistic expertise to create a pronunciation dictionary. As the number of supported languages continuously grows, it will dramatically increase the effort required to train, deploy, and maintain so many ASR systems in a production environment. Moreover, the code-switching phenomenon [1] that contains more than one language within an utterance is another great challenge to ASR service. How to deal with these multilingual and code-switching problems have gained more and more attention.

Previous works on multilingual speech recognition mostly focused on making the acoustic model multilingual, which can be divided into two categories: i) acoustic modeling with an universal phone set [2, 3, 4, 5]; ii) acoustic modeling with a multilingual-DNN [6, 7, 8, 9, 10, 11]. As to multilingual-DNN, the lower layers are shared across languages and the output layer is language-specific. Consequently, it allows a shared, language-independent speech representation to be more robustly learned in the lower layers due to the increased training data presented, which is especially helpful for low-resource speech recognition. However, it still needs language-specific prior and posterior probabilities during model inference.

Recently, researchers have been interested in building multilingual ASR system in the so-called end-to-end framework. In [12], the end-to-end multilingual system is trained using CTC-based approach [13] with an universal character set as modeling units. Experimental results show that it can outperform both individual monolingual systems and systems built with the multilingual-DNN approach. [14] demonstrates that LAS model [15], which takes a union of language-specific grapheme sets as modeling units and jointly trained across data from 9 Indian languages without any explicit language specification, consistently outperforms monolingual LAS models trained independently on each language. Unfortunately, as analyzed in [14], LAS model is unable to perform code-switching due to the issue with attention-based sequence-to-sequence models that the language model is dominating the acoustic model [16].

Code-switching speech recognition presents great challenges in acoustic, language and pronunciation modeling since getting enough data at code-switched points for both the acoustic model and the language model is arduous. Accordingly, previous works have proposed the language identification (LID) based approach [17], acoustic modeling based approach [18, 19] and language modeling based approach [20, 21] to deal with this problem. Recently, researchers have proposed to improve the performance of code-switching speech recognition system by using the popular end-to-end approach [22, 23, 24]. The advantage is that it can jointly optimize the acoustic model and the language model without the need of expert linguistic knowledge.

Previous works usually considered multilingual and code-switching speech recognition as two isolated tasks. As a result, a multilingual ASR system is unable to deal with the code-switching problem while a code-switching ASR system can not match the performance of a language-specific ASR system on monolingual cases. In this work, we study how to build a language-universal ASR system that can recognize not only monolingual speech but also code-switching speech without any language-specific information during model inference. Particularly, we showcase the capability of our model on the bilingual Mandarin-English speech recognition. Firstly, we have constructed a set of universal acoustic modeling units, namely *Character-Subword*, which adopts Chinese characters for Mandarin and Byte Pair Encoding (BPE) [25] based subwords for English. Secondly, we have proposed a novel bilingual acoustic model that consists of a shared output layer corresponding to the Character-Subword acoustic modeling units and two language-specific subnets. Moreover, the well-trained DFSMN-CTC-sMBR [26] based monolingual Mandarin and English acoustic models are used to initialize these subnets. And then, the bilingual acoustic model is optimized using both the Mandarin and English training data with the CTC criterion followed by state-

level minimum Bayes risk (sMBR) based sequence-level discriminative training. Our intention behind the proposed bilingual acoustic model is opposite to multilingual-DNN [6, 7, 8]. We want the bilingual model to have strong language discrimination ability instead of to learn a language-independent shared representation. Thirdly, a mixed language model is built by interpolating the trigram Mandarin and English language models. Finally, we have evaluated the proposed language-universal ASR approach on three test sets, namely Mandarin-only, English-only and Mandarin-English code-switching test sets.

2. Monolingual ASR system

Recently, neural networks based acoustic modeling have become the mainstream. There exist various types of neural network architectures. In this section, we will give a briefly review on the *DFSMN-CTC-sMBR* [26] acoustic model.

2.1. DFSMN

DFSMN [27] is an improved *Feed-forward Sequential Memory Networks* (FSMN) [28] structure that builds a very deep architecture by introducing skip connections. The key element in DFSMN is the so-called DFSMN-component, which enables DFSMN to model the long-term dependency in sequential signals while without using recurrent feedback. The DFSMN-component consists of four parts: ReLU layer, linear projection layer, memory block and skip connection. The operations of memory block in ℓ-th DFSMN component take the following form:

$$\mathbf{m}_t^\ell = \mathbf{m}_t^{\ell-1} + \mathbf{p}_t^\ell + \sum_{i=0}^{N_1^\ell} \mathbf{a}_i^\ell \odot \mathbf{p}_{t-s_1*i}^\ell + \sum_{j=1}^{N_2^\ell} \mathbf{c}_j^\ell \odot \mathbf{p}_{t+s_2*j}^\ell \quad (1)$$

Here, $\mathbf{p}_t^\ell$ denote the outputs of the linear projection layer and $\mathbf{m}_t^\ell$ denotes the output of the memory block. N_1^ℓ and N_2^ℓ denote the look-back order and lookahead order of the memory block, respectively. s_1 is the stride factor of look-back filter and s_2 is the stride of lookahead filter.

2.2. CTC-SMBR

Connectionist temporal classification (CTC) [29] is a loss function for sequence labeling problems, which converts the sequence of labels with timing information into the shorter sequence of labels by removing timing and alignment information. The main idea is to introduce the additional CTC blank ($-$) label during training, and then remove the blank labels and merge repeating labels to obtain the unique corresponding sequence during decoding.

For a set of target labels, Ω, and its extended CTC target set is defined as $\bar{\Omega} = \Omega \cup \{-\}$. Given an input sequence $\mathbf{x}$ and its corresponding output label sequence $\mathbf{y}$. The CTC path, π, is defined as a sequence over $\bar{\Omega}$, $\pi \in \bar{\Omega}^T$, where T is the length of the input sequence $\mathbf{x}$. The label sequence $\mathbf{y}$ can be represented by a set of all possible CTC paths, $\Phi(\mathbf{y})$, that are mapped to $\mathbf{y}$ with a sequence to sequence mapping function $\mathcal{F}$, $\mathbf{y} = \mathcal{F}(\mathbf{\Phi}(\mathbf{y}))$. Thereby, the log-likelihood of reference label sequence $\mathbf{y}$ given the input $\mathbf{x}$ can be calculated as an aggregation of the probabilities of all possible CTC paths:

$$p(\mathbf{y}|\mathbf{x}) = \sum_{\pi \in \Phi(\mathbf{y})} p(\pi|\mathbf{x}) \quad (2)$$

Model training can be carried out by minimizing the negative log-likelihood.

CTC is still a frame-wise discriminative training criterion, which is suboptimal for word error rate minimization objective in ASR. Thereby, it can be further optimized with sequence-level discriminative training criteria such as state-level minimum Bayes risk (sMBR) criterion [30].

3. Language-universal ASR system

In this section, we introduce our approach to construct a language-universal Mandarin-English ASR system. Here, *language-universal* means the model that is not explicitly given any language-specific information should be able to recognize Mandarin, English and Mandarin-English code-switching speech. we have explored acoustic modeling units, bilingual acoustic model and inference method in order to move towards this goal.

3.1. Acoustic modeling units

In [26], it has investigated the performance of DFSMN-CTC-sMBR acoustic models with context-independent Initial-Final (CI-IF), context-dependent Initial-Final (CD-IF), syllable and a novel mixed Character-Syllable as modeling units for Mandarin speech recognition. Experimental results suggest that the *hybrid Character-Syllable* modeling units, which mix the high frequency Chinese characters and syllables, is the best choice for Mandarin speech recognition. For *hybrid Character-Syllable*, the low frequency characters are mapped into the syllables to deal with the OOV problem. In this work, instead of mapping the low frequency characters into syllables, we propose to map the low frequency characters into the high frequency characters with the same pronunciation. As a result, we come up with a pure Chinese characters based modeling units for Mandarin speech recognition. Specifically, we keep the top 2000 Chinese characters as acoustic modeling units and map the other characters into these top 2000 characters. As to English, the widely used acoustic modeling units in end-to-end speech recognition systems are characters [15, 31] and subwords [32]. In this work, we adopt the BPE [25] as subword segmentation method to generate 1000 subwords as acoustic modeling units for English.

For language-universal Mandarin-English system, we combine the Chinese characters and English subwords to form an universal modeling units set, namely *Character-Subword*. The transcriptions can be converted into the Character-Subword label sequences for CTC-based model training.

3.2. Baseline bilingual acoustic model

Previous work in [12] proposed a CTC-based bilingual end-to-end model for English-Spanish speech recognition, which is a common CTC-based acoustic model with an universal character set as modeling units trained using the bilingual training data. Experimental results show that the bilingual models perform as well as the monolingual models even if no language identification information is provided to the network. In this work, we adopt this method to train the baseline bilingual Mandarin-English model. We use the *Character-Subword* as modeling units, DFSMN-CTC-sMBR as acoustic model and a large bilingual Mandarin-English corpus as training data.

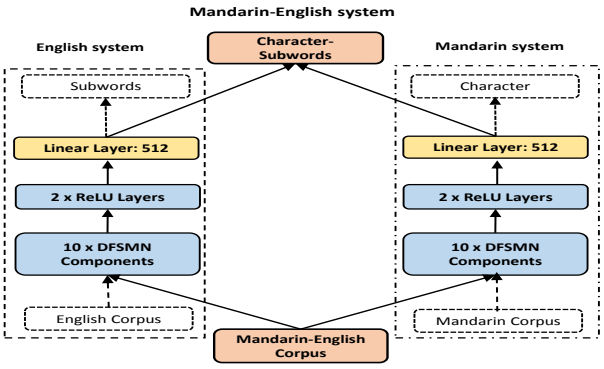

Figure 1: *Bilingual Mandarin-English acoustic model.*

3.3. Novel bilingual acoustic model

The widely used acoustic model in previous multilingual works [6, 7, 8, 9] is the multilingual-DNN, in which the hidden layers are shared across many languages while the softmax layers are language-specific. The shared hidden layer can be considered as an universal feature transformation [8], which can be transferred to boost the performance of other languages, especially in languages with limited resources. Unfortunately, this type of multilingual system is language-dependent during model inference. Moreover, we doubt that this shared hidden layer architecture maybe not suitable for resource-rich multilingual speech recognition, such as Mandarin-English. If the output of shared hidden layer does not have strong language discrimination, the language-specific softmax output layers will not have enough distinguishing ability. Moreover, if we want the multilingual model to be able to deal with the code-switching problem, the acoustic model needs to have strong language discrimination ability. Based on these intuitions, we propose a novel bilingual acoustic model, as shown in Figure 1, which consists of two separated subnets and a shared output layer outputting the posterior probabilities for *Character-Subword* acoustic modeling units. The well-trained CTC based monolingual English and Mandarin systems are used to initialize these two subnets in the bilingual model. This initialization is very important, since it enables the bilingual model to have strong language distinguishing ability from the beginning of model training. We fine-tune this model using the mixed monolingual Mandarin and English data with CTC and sMBR criterion.

3.4. Model inference

For model inference, we use the standard beam search with weighted finite-state transducers [33, 34]. The search graph is built by composing the WFST of language model *G*, lexicon *L* and CTC token *T*. Depending on the resource used to build the search graph, model inference can be divided into two categories, language-dependent and language-independent decoding. For language-dependent decoding, the language-specific language model, lexicon and prior are used during model inference. As to language-independent decoding, we don't know speech that need to be decoded belong to Mandarin, English or Chinese-English code-switching. Therefore, we firstly build a mixed language model by interpolating the N-gram Mandarin language model and English language model. And then, the mixed language model and bilingual lexicon are used to build the search graph. During language-independent model inference, we will not provide any language-specific information.

Table 1: *Performance of the baseline monolingual ASR systems.*

System	Criterion	Mand.(CER%)	Eng.(WER%)
Mandarin	CTC	8.08	-
	+SMBR	**6.98**	-
English	CTC	-	13.47
	+SMBR	-	**11.31**

Table 2: *Performance of the baseline bilingual systems decoded with (w) and without (w/o) language-specific information.*

ID	Data		Lang. Info.	Mand. (CER%)	Eng. (WER%)
	Mandarin	English			
1	100%	10%	w	8.28	22.61
			w/o	8.44	23.23
2	100%	50%	w	9.08	16.44
			w/o	9.60	17.01
3	100%	100%	w	**9.52**	**15.04**
			w/o	9.83	15.66

4. Experiments

4.1. Experimental setup

We conduct our experiments on two large Mandarin and English corpora that consists of about 20,000 hours and 15,000 hours data respectively. We divide the data into training set and development set, which contain of 97% and 3% data, respectively. These monolingual Mandarin corpus and English corpus are further mixed to form the bilingual Mandarin-English corpus, which is used to train bilingual acoustic models. For model inference, we have built three test sets: Mandarin test set (about 30 hours), English test set (about 10 hours) and Mandarin-English code-switching test set (about 5 hours). Acoustic feature used for all experiments are 80-dimensional log-mel filterbank (FBK) energies computed on 25ms window with 10ms shift. We stack the consecutive frames within a context window of 5 (2+1+2) to produce the 400-dimensional features and then down-sample the inputs frame rate to 30ms. During decoding, pruned trigram Mandarin and English language models are used for language-dependent Mandarin and English decoding respectively. The mixed language model, which is generated by interpolating the trigram Mandarin and English language models, is used for language-independent decoding.

4.2. Monolingual ASR system

For the baseline monolingual ASR systems, we have trained *DFSMN-CTC-sMBR* acoustic models for Mandarin and English respectively. The model topology of DFSMN is denoted as: $5 * 80\text{-}N_f \times [2048\text{-}512(N_1; N_2; s_1; s_2)]\text{-}2 \times 2048\text{-}512\text{-}N_{out}$. Here, $N_f = 10$, $N_1 = 5$, $N_2 = 2$, $S_1 = 2$, $S_2 = 1$. N_{out} denotes the output layer size which is equal to the number of acoustic modeling units plus 1 (denote the *blank*). These models are first trained with the CTC loss and then further optimized by the sMBR based sequence discriminative training. During model inference, the language-dependent decoding is adopted for these baseline monolingual models. Detailed experimental results are as shown in Table 1.

4.3. Bilingual ASR system

We first build the baseline Mandarin-English bilingual systems using the *DFSMN-CTC* acoustic models (without sMBR train-

Table 3: *Performance of the proposed bilingual systems decoded with(w) and without(w/o) language-specific information.*

Lang. Info.	Criterion	Mand.(CER%)	Eng.(WER%)
w	CTC	8.04	12.80
	+SMBR	**6.94**	**11.33**
w/o	CTC	8.14	12.94
	+SMBR	**7.02**	**11.60**

Table 4: *Comparison of the baseline and the proposed bilingual systems on code-switching test set. (M-Mandarin; E-English)*

System	LM	Code-Switching Test Set		
		M	E	All
Baseline Bilingual	3-gram	8.73	20.74	**9.51**
	1-gram	10.14	21.71	10.89
Proposed Bilingual	3-gram	7.70	17.03	**8.31**
	1-gram	9.03	17.82	9.60

ing) with *Character-Subword* as modeling units, as introduced in Sec.3.2. We have investigated the influence of Mandarin-English data mixing ratio: 1:0.1, 1:0.5 and 1:1. Here, 1:0.1 means we use the whole Mandarin data and 10% English data. Decoding of these bilingual systems can be language-dependent or language-independent, according to whether is given the language-specific information. Experimental results in Table 2 shown that this kind bilingual system, which is not explicitly given any information about language identity, is able to recognize both Mandarin and English. We can further improve the performance by using the language-dependent decoding. However, the performance of Mandarin recognition is declining while the performance of English recognition is improving with the increasing of English training data. Compared to the monolingual Mandarin and English systems (in Table 1), the baseline Mandarin-English bilingual system trained with the whole bilingual data will suffer from more than 10% performance degradation in both languages.

For the proposed bilingual system, the model topology is as shown in Figure 1. The baseline monolingual Mandarin and English acoustic models in Sec.4.2 are used to initialize the bilingual acoustic model (AM). This bilingual AM is first trained with the CTC loss and then further optimized by the sMBR based sequence discriminative training using the Mandarin-English bilingual data. Similar to the baseline bilingual experiments, we have evaluated the proposed bilingual AM based Mandarin-English system on the Mandarin and English test sets under language-dependent and language-independent model inference setups. Experimental results in Table 3 show that the proposed bilingual system can achieve similar performance whether decoded with or without language-specific information. Moreover, compared to the baseline monolingual systems in Table 1, it can achieve comparable performance in both Mandarin and English test sets, without given any information about language identity.

In Table 4, we have compared the performance of the baseline and the proposed bilingual systems on a Mandarin-English code-switching test set. The code-switching test set contains 3300 manually recorded voices. For the baseline bilingual system we choose the *ID3* model (in Table 2) and then continue to optimize it using sMBR criterion. Performance is evaluated in term of character error rate (CER) for Mandarin, word error rate (WER) for English and Mixed error rate (MER) for all test set.

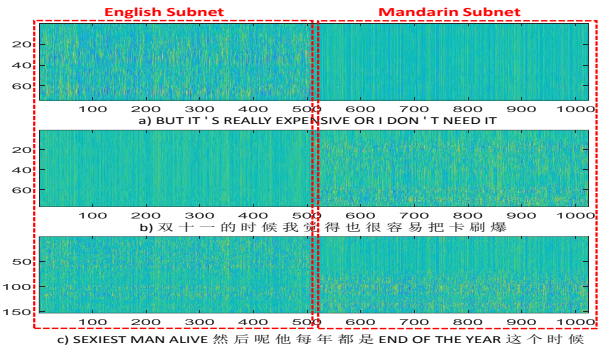

Figure 2: *Examples of discriminative features (from the linear layer of Figure 1) learned by the proposed bilingual Mandarin-English acoustic model.(The vertical axis denotes the index of speech frame.)*

Results in the second and fourth rows of Table 4 show that the proposed method can achieve 11.8% and 17.9% relative error reduction on Mandarin and English parts, respectively. We further investigate the effect of language model to the performance of the bilingual systems. For both the baseline and proposed bilingual systems, the performance of the Mandarin part suffer from more obvious performance degradation than English part when switching from 3-gram to 1-gram. This indicates that the language model plays a limited role in the recognition of English words in Mandarin-English code-switching sentences. As to further improve the performance of code-switching systems, we may pay attention to improving language model coverage for code-switching or enhancing the ability of acoustic model to distinguish between languages. In Figure 2, we have plotted the spliced outputs of the linear layers (yellow marked modules in Figure 1) in the proposed bilingual acoustic model for three sentences. It seems that the proposed bilingual AM can automatically learn the language distinguishing ability, where the English subnet can generate discriminative features for English inputs while almost uniform distributed feature for Mandarin inputs and vice versa. Moreover, the bilingual acoustic model can automatically learn code-switching as shown in Figure 2c.

5. Conclusions

In this paper we proposed universal bilingual system for Mandarin-English speech recognition. Specifically, we come up with a novel bilingual acoustic model that consists of two monolingual system initialized subnets and a shared output layer corresponding to the *Character-Subword* acoustic modeling units. When trained on a large bilingual Mandarin-English corpus, the proposed bilingual acoustic model can automatically learn language distinguishing ability. Experimental results show that this model, which is not explicitly given any information about language identity, can achieve comparable performance in recognizing monolingual speech (Mandarin or English) when compared to the baseline monolingual systems. Moreover, the proposed bilingual model has demonstrated the great potential to deal with the code-switching problem. Compared to a baseline bilingual system, the proposed language-universal Mandarin-English system can achieve 11.8% and 17.9% relative error reduction on Mandarin and English parts of a code-switching test set, respectively. For further work, we may generalize this type of language-universal ASR system to more language combinations.

6. References

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

[2] T. Schultz and A. Waibel, "Fast bootstrapping of LVCSR systems with multilingual phoneme sets," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[3] 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 *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4333–4336.

[4] N. T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, and H. Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7639–7643.

[5] B. Li, Y. Zhang, T. Sainath, Y. Wu, and W. Chan, "Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes," *arXiv preprint arXiv:1811.09021*, 2018.

[6] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8619–8623.

[7] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7319–7323.

[8] 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 *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7304–7308.

[9] D. Chen and B. K.-W. Mak, "Multitask learning of deep neural networks for low-resource speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 7, pp. 1172–1183, 2015.

[10] T. Sercu, C. Puhrsch, B. Kingsbury, and Y. LeCun, "Very deep multilingual convolutional neural networks for LVCSR," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4955–4959.

[11] S. Tong, P. N. Garner, and H. Bourlard, "An investigation of deep neural networks for multilingual speech recognition training and adaptation," in *Annual Conference of the International Speech Communication Association*. IEEE, 2017.

[12] S. Kim and M. L. Seltzer, "Towards language-universal end-to-end speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4914–4918.

[13] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning*, 2014, pp. 1764–1772.

[14] 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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4904–4908.

[15] 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.

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

[17] J. Weiner, N. T. Vu, D. Telaar, F. Metze, T. Schultz, D.-C. Lyu, E.-S. Chng, and H. Li, "Integration of language identification into a recognition system for spoken conversations containing code-switches," in *Spoken Language Technologies for Under-Resourced Languages*, 2012.

[18] 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 *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4889–4892.

[19] 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.

[20] H. Adel, N. T. Vu, K. Kirchhoff, D. Telaar, and T. Schultz, "Syntactic and semantic features for code-switching factored language models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 431–440, 2015.

[21] G. I. Winata, A. Madotto, C.-S. Wu, and P. Fung, "Learn to code-switch: Data augmentation using copy mechanism on language modeling," *arXiv preprint arXiv:1810.10254*, 2018.

[22] 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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4919–4923.

[23] N. Luo, D. Jiang, S. Zhao, C. Gong, W. Zou, and X. Li, "Towards end-to-end code-switching speech recognition," *arXiv preprint arXiv:1810.13091*, 2018.

[24] 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.

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

[26] S. Zhang, M. Lei, Y. Liu, and W. Li, "Investigation of modeling units for mandarin speech recognition using DFSMN-CTC-SMBR," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[27] S. Zhang, M. Lei, Z. Yan, and L. Dai, "Deep-FSMN for large vocabulary continuous speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5869–5873.

[28] S. Zhang, C. Liu, H. Jiang, S. Wei, L. Dai, and Y. Hu, "Feed-forward sequential memory networks: A new structure to learn long-term dependency," *arXiv preprint arXiv:1512.08301*, 2015.

[29] 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*. ACM, 2006, pp. 369–376.

[30] H. Sak, F. de Chaumont Quitry, T. Sainath, K. Rao *et al.*, "Acoustic modelling with cd-ctc-smbr lstm rnns," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 604–609.

[31] J. Li, G. Ye, R. Zhao, J. Droppo, and Y. Gong, "Acoustic-to-word model without OOV," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 111–117.

[32] T. Zenkel, R. Sanabria, F. Metze, and A. Waibel, "Subword and crossword units for ctc acoustic models," *arXiv preprint arXiv:1712.06855*, 2017.

[33] H. Sak, A. Senior, K. Rao, O. Irsoy, A. Graves, F. Beaufays, and J. Schalkwyk, "Learning acoustic frame labeling for speech recognition with recurrent neural networks," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 4280–4284.

[34] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 167–174.