Multi-lingual Transformer Training for Khmer Automatic Speech Recognition

Kak Soky*, Sheng Li†, Tatsuya Kawahara‡ and Sopheap Seng*
* National Institute of Posts, Telecoms and ICT (NIPTICT), Phnom Penh, Cambodia
E-mail:soky.kak@niptict.edu.kh and sopheap.seng@niptict.edu.kh
† National Institute of Information and Communications Technology (NICT), Kyoto, Japan
E-mail: sheng.li@nict.go.jp
‡ Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto, Japan
E-mail: kawahara@i.kyoto-u.ac.jp

***Abstract*—Currently, there are three challenges for constructing reliable ASR systems for the Khmer language: (1) the lack of language resources (text and speech corpora) in digital form, (2) the writing system without explicit word boundary, and (3) the pronunciation model is not well studied. In this paper, to avoid the extensive work on selecting proper acoustic units (e.g., phones, syllables) and preparing the frame-level labels on the traditional DNN-HMM framework, we directly use words or characters as the label using state-of-the-art transformer-based end-to-end model. Moreover, we use the multi-lingual training framework to tackle the low-resource data problem. All experiments are performed on the Basic Expressions Travel Corpus (BTEC) datasets. The experiments show that the proposed multi-lingual transformer-based end-to-end model can achieve significant improvement compared to the DNN-HMM baseline model[1].**

I. Introduction

As the most natural way of communication, voice interface with the support of the automatic speech recognition (ASR) technology has become crucial in human-computer interaction (HCI) in various devices of today's digitized society. At this moment, most commercial ASR-enabled products focus on popular languages such as English, French, Chinese, and Japanese. As the growth and the internationalization of the ASEAN region, the speech recognition of ASEAN languages such as the Khmer language has become a topic worthy of extensive research.

Conventional automatic speech recognition (ASR) systems, both GMM-HMM [1] and DNN-HMM [2], require huge amounts of resources. However, the Khmer language is one of the under-resourced southeast Asian languages that bears the big challenging characteristics: (1) the lack of language resources (text and speech corpora) in digital form, (2) the writing system without explicit word boundary, and (3) the pronunciation model is not yet well studied. Thus, it is necessary to study efficient acoustic modeling techniques for this language.

The end-to-end neural network model simplifies the ASR system construction, and solves the sequence labeling problem between variable-length speech frame inputs and label outputs (phone, character, syllable, word, etc.), and achieved promising results on many ASR tasks. Various types of end-to-end model have been studied in recent years, i.e. connectionist temporal classification (CTC) [3], [4], attention-based encoder-decoder (Attention) end-to-end models [5], [6], and their variants [7], [8], [9], [10], [11]. Recently, the transformer [12] has achieved promising results. The transformer-based model maps an input speech feature sequence to a sequence of intermediate representations in the encoder. Then, the decoder generates an output sequence of symbols (phonemes, syllables, words, subwords, or words) given the intermediate representations. The most significant difference with those commonly used end-to-end models [5], [6] is that the transformer-based acoustic model relies on no-recurrence components [12].

In this paper, we investigate transformer-based end-to-end model using character-based units. This work focuses on tackling the low-resource data problems using the multi-lingual training framework, which is first time evaluated by the transformer-based models to our best knowledge. All experiments are performed on the Basic Expressions Travel Corpus (BTEC) [13] datasets.

The rest of this paper is organized as follows. The related works are overviewed in Section II. In Section III, the data sets and the baseline systems of this paper are introduced. In Section IV, the proposed method for our task is explained and evaluated. This paper concludes in Section V.

II. Related Work

Khmer ASR has been investigated with GMM-HMM-based techniques [14], which evaluated grapheme-based vs. phoneme-based acoustic model, word and sub-word language models, hybrid word/sub-word language models, and ASR outputs combination by using simple N-best list voting mechanism. In [15], an LVCSR system was developed on the broadcast news transcription on the Khmer language. They experimented on three different language models (LMs): word based LM, syllable-based LM, and character cluster (CC) based LM in grapheme context independent and dependent manners. Both of these systems could not give good performance for real applications, but they provide many useful tips and language processing tools for the Khmer language.

[1]The work was performed during Mr. Kak Soky was in NIPTICT. He is currently with Ministry of Education, Youth, and Sports (MoEYS), Cambodia. Email: kak.soky.hs@moeys.gov.kh

III. Data Description and Baseline Systems

A. Writing System of Khmer

There are no standard rules for using spaces in the Khmer writing system. Spaces are not used to separate words in the writing style, but occasionally used to be easier in reading. The large contiguous blocks of unsegmented words cause major problems for natural language processing applications such as machine translation, speech recognition, speech synthesis, and information extraction. Moreover, word segmentation is not a trivial task. Along with those problems, we found four types of segmentation of Khmer text have been proposed to break a Khmer sentence in previous studies.

Sentence	ភាសាខ្មែរមានភាពស្មុគស្មាញ
Word	ភាសា ខ្មែរ មាន ភាព ស្មុគស្មាញ
Syllable	ភា សា ខ្មែ រ មាន ភាព ស្មុ គ ស្មា ញ
Character cluster	ភា សា ខ្មែ រ មា ន ភា ព ស្មុ គ ស្មា ញ
Character	ភ ា ស ា ខ ្ម ែ រ ម ា ន ភ ា ព ស ្ម ុ គ ស ្ម ា ញ

Fig. 1. Segmentation of Khmer scripts

The first is character cluster (CC) or unbreakable unit that is the smallest unit of Khmer words, but it is just mono-syllable, meaningless, and sometimes not readable. This CC form has been used to improve Khmer word segmentation [16], [17], and for building Khmer ASR [14].

The second one is syllable, which is a unit of pronunciation having one or more consonant sounds, accompanied with or without surrounding vowels. It forms the whole or a part of a word, which is readable, but sometimes meaningless. Syllable segmentation was built in [15] by applying 20 linguistic rules on mono-syllable for a Khmer ASR system.

The third is word, which is a meaningful element of speech or writing, and can be mono-syllable or multiple syllables. In [15], words are used for a Khmer ASR system. In addition, [18] defines four classes of Khmer word: single words, compound words, compound words with prefix, and compound words with suffix. Single word is the smallest unit of word that has one or more syllables and meaningful. Compound word is a word composed of two or more single words. They used conditional random fields (CRF) to segment Khmer words into two types of words. Compound word is annotated by special tokens, such as $\sim$, to connect two or more single words. This tool achieves an average F-score is 0.99 on 12,468-sentence test set. In this study, we will use word segmentation tool [18] to segment a Khmer text corpus by breaking down a compound word into single words, which will be used to build a lexicon dictionary, LM, and AM transcription.

The fourth one is character, which is a simplification of CC. It has a smaller number of units. In Section IV, we use this character unit to compare with the single word unit for end-to-end models.

B. Speech Corpus

The collection of audio data and transcription of that data in a large amount is expensive and time-consuming. For these reasons, we currently limit our ASR system to the travel domain for which NICT has developed a set of parallel multi-lingual text corpora the Basic Travel Expression Corpus (BTEC) [13] available for many languages including Khmer. This Khmer language corpus contains 96K unique sentences with wide coverage of expressions in the traveling domain. In this study, this Khmer BTEC data will be used as the resource for both speech and textual data.

The recording was conducted in three places as shown in Table I.

TABLE I
Enviroment of Speech Data Collection

Place	Device	#Speaker	#Utterances
Sound-proof room (NICT)	Microphone	15M+20F	37,800
Office (NIPTICT)	Smartphone+Headset	3M+9F	10,569
Public (Cambodia)	Smartphone	16M+15F	12,551
/	/	34M+44F	60,920

Finally, a total of 60,920 utterances, about 107 hours (including silence) of a spoken corpus is constructed for acoustic model training. For the testing set, we prepare two sets: Open test-set (Open) is about 5.5-hour long. Five speakers are selected from the public place speech corpus.

C. Lexicon

There are 34K words collected from Chuon Nat dictionary [19], Khmer websites, and some other new words appeared in BTEC data. We first selected 10K words to transcribe manually as a pronunciation dictionary. Then, we built a statistical grapheme-to-phoneme (G2P) based on a weighted finite state transducer (WFST). After that we have a model that can generate pronunciation entries for the other words. The generated result from the WFST was manually checked. Finally, we built a lexicon of 34K words.

D. Language Model (LM)

We built a 3-gram LM by using SRILM toolkit [20] and interpolated this model with the Witten-Bell smoothing algorithm. The training text consists of about 96K sentences of BTEC data that was cleaned and segmented by using Khmer word segmenation tool [18].

E. Baseline ASR Systems

Table II shows the settings for the baseline DNN models. Sigmoid hidden units and softmax output units were used. The initial learning rate was set to 0.008, and the minibatch size to 256. The input to the network was an 11-frame context window (the current frame and five frames on each side of the current frame). Two kinds of DNN models are explored in this system. One is the DNN trained using the cross-entropy criterion (DNN-CE), and the other is the DNN based on sequence discriminative training using state-level minimum Bayers risk criterion (DNN-sMBR). DNNs are trained directly

on top of the GMM-bMMI model, using 11 frames of context windows of the fMLLR features. For parameters of the DNNs, 418 units of the input layer, five hidden layers, and 512 units in each layer are used. The output unit number of the DNNs is 3907. All the acoustic model training steps are conducted by using Kaldi speech recognition toolkit [21], and the DNN-baseds models are trained on a single GPU (tesla-K40).

TABLE II
SUMMARY OF ACOUSTIC MODEL TRAINING

	Settings
Feature	LDA-MLLT-SAT
Splicing	11 frames (5 left + 1 center + 5 right)
States	3,907
Layers	418x(512x5)x3907

The performance of a GMM and two DNNs is shown in Table III on the open test-set (Open). We used the NICT SprinTra decoder [22] for decoding. The recognition performance of DNN-CE and DNN-sMBR significantly outperform the GMM-bMMI by 3.3% and 3.8% absolute in CER%.

TABLE III
CHARACTER ERROR RATE (CER%) OF BASELINE MODELS

Baseline Models	CER% on (Open)
GMM-bMMI	7.0%
DNN-CE	3.7%
DNN-sMBR	3.2%

IV. PROPOSED END-TO-END ASR SYSTEM

In order to solve the problems of lacking resources, we build the end-to-end attention-based models that integrate the acoustic, pronunciation and language models into a single neural network. In this paper, we used the transformer-based neural machine translation (NMT)[12] in tensor2tensor[2] for all our experiments.

As shown in Figure 2, we jointly train four low-resourced South/Southeast Asian languages (Myanmar, Khmer, Sinhalese and Nepalese) to solve the low-resource problem. When training the multilingual model, we add the particular token <Language Mark> (e.g., <MY>, <KH>, <SI> and <NE>) to the beginning of the utterances. The training labels are organized as "<S> <Language Mark> labels </S>". The knowledge of language identification is provided by these <Language Mark>s. They will be first recognised out and guide next steps of decoding. This is an efficient multilingual learning for Transformer-based ASR, which is different from conventional multitask DNN training.

The Khmer data set is the BTEC data set which is the same as the baseline in the previous section. The Myanmar data set is the same as our previous work [23]. The data sets of the other two languages are selected from Google's open-source database[3]. These datasets are all from smartphone

[2]https://github.com/tensorflow/tensor2tensor
[3]http://www.openslr.org/52/ and http://www.openslr.org/54/

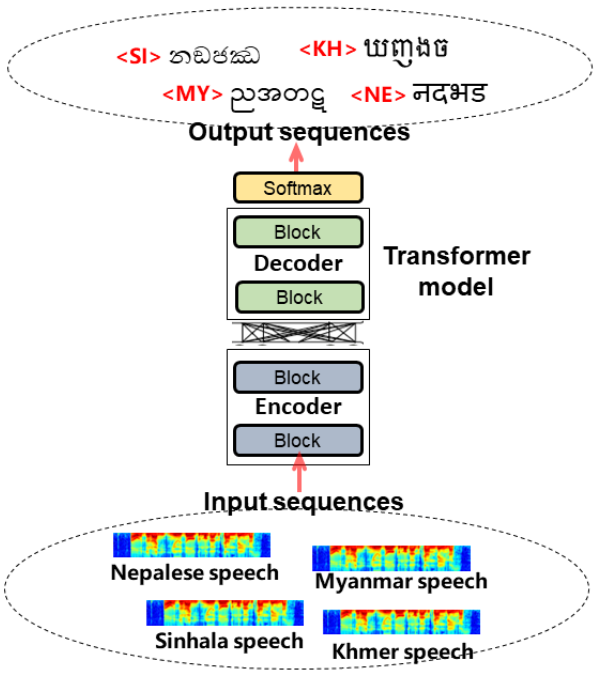

Fig. 2. Multi-lingual Transformer Training for Khmer Automatic Speech Recognition

input in tourist scenarios. For each language, we select speech data for training, as shown in Table IV. The quality of the Khmer speech data is the best. The number of the words and characters in each individual dataset and their mixture are listed in Table V.

TABLE IV
MULTILINGUAL DATASETS

Language	Dataset	Hours
Myanmar (**MY**)	Training	54.4
Sinhalese (**SI**)	Training	27.9
Nepalese (**NE**)	Training	38.7
Khmer (**KH BTEC**)	Training	102.2
	Testing (Open)	5.5

TABLE V
NUMBER OF DIFFERENT MODELING UNITS

Language	**#word**	**#character**
Myanmar (**MY**)	17,951	67
Khmer (**KH**)	3,129	79
Sinhalese (**SI**)	24,803	153
Nepalese (**NE**)	25,929	100
Multilingual (**MY+KH+SI+NE**)	71,812	365

We used 120-dim filterbank features (40-dim static +Δ +$\Delta\Delta$), which were mean and variance normalized per speaker, and four frames were spliced (four left, one current and zero right). The training and testing settings are listed in Table VI.

From table VII, we can conclude that the transformer-based

TABLE VI
MAJOR EXPERIMENTAL SETTINGS

Model structure			
Attention-heads	8	Decoder-blocks	6
Hidden-units	512	Residual-drop	0.3
Encoder-blocks	6	Attention-drop	0.0
Training settings			
Max-length	5000	GPUs (K40m)	4
Tokens/batch	10000	Warmup-steps	12000
Epochs	30	Steps	300000
Label-smooth	0.1	Optimizer	Adam
Testing settings			
Ave. chkpoints	last 20	Batch-size	100
Length-penalty	0.6	Beam-size	13
Max-length	200	GPUs (K40m)	4

TABLE VII
ASR PERFORMANCE (CER%) OF ACOUSTIC MODELS WITH DIFFERENT SETTINGS ("**W**" MEANS WORD-BASED MODEL, "**C**" MEANS CHARACTER-BASED MODEL)

Models	#Acoustic Units	Evaluations (Openset)
DNN-sMBR	3,907	3.2%
single-language (**w**)	3,129	1.3%
single-language (**c**)	79	5.0%
multi-lingual (**w**)	71,812	1.5%
multi-lingual (**c**)	365	**0.7%**

The results compared to the lowest result without statistical significance (from two-tailed t-test at significant level of p-value < 0.05) are shown in bold fonts.

end-to-end model achieved the best significant improvement compared with the DNN baseline system. Moreover, multi-lingual training can effectively enhance the Khmer end-to-end ASR system. In single language setting, the word-based model is better than the char-model and the baseline because it has lexical constraint, while the char-based model is worse than the baseline as it does not use a lexicon and LM. In multi-lingual setting, the char-based model benefits more from joint training as characters share acoustic characterisitcs (some characters are shared among these languages as shown in Table V), while the word-based model does not benefit because word units are mostly disjoint in these languages.

V. CONCLUSIONS

This paper presents an LVCSR system for the Khmer language using the current state-of-the-art transformer-based end-to-end modeling technology. This paper focuses on tackling the low-resource data problems using multi-lingual training framework. Experiments show that the multilingual training using character units greatly improved ASR performance.

ACKNOWLEDGMENT

We would like to express our gratitude to Dr. Hiroaki Kato and Dr. Xugang Lu for their supervision and facilitation in this work. We are thankful to Chuon Vanna, Sorn Kea, Leng Sreyhuy of Research and Innovation Center, National Institute of Posts, Telecoms and ICT, Cambodia for their help recording and segmentation speech data for this work and we are also grateful to Hour Kaing and colleagues in NICT for their suggestions, discussions, and support in building this Khmer LVCSR system.

REFERENCES

[1] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1988.
[2] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE Trans. ASLP*, vol. 20, no. 1, pp. 30–42, 2012.
[3] A. Graves and N. Jaitly, "Towards End-to-End speech recognition with recurrent neural networks," in *Proc. ICML*, 2014.
[4] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-End speech recognition using deep RNN models and WFST-based decoding," in *Proc. IEEE-ASRU*, 2015, pp. 167–174.
[5] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. NIPS*, 2015.
[6] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE-ICASSP*, 2016.
[7] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "End-to-end speech recognition using lattice-free mmi," in *Proc. INTERSPEECH*, 2018.
[8] 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.
[9] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "Espnet: End-to-end speech processing toolkit," in *Proc. INTERSPEECH*, 2018.
[10] S.Ueno, H.Inaguma, M.Mimura, and T.Kawahara, "Acoustic-to-word attention-based model complemented with character-level ctc-based model," in *Proc. IEEE-ICASSP*, 2018.
[11] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, "Advances in joint CTC-Attention based End-to-End speech recognition with a deep CNN Encoder and RNN-LM," in *Proc. INTERSPEECH*, 2017.
[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *arXiv preprint arxiv:1706.03762*, 2017.
[13] G. Kikui, E. Sumita, T. Takezawa, and Seiichi Yamamoto, "Creating corpora for speech-to-speech translation," in *Proc. EUROSPEECH*, 2003.
[14] S. Seng, S. Sam, V.-B. Le, B. Bigi, and L. Besacier, "Which units for acoustic and language modeling for khmer automatic speech recognition," in *Proc. International Workshop on Spoken Languages Technologies for Under-Ressourced Languages*, 2008.
[15] S. Seng, S. Sam, L. Besacier, B. Bigi, and E. Castelli, "First broadcast news transcription system for khmer language," in *Proc. Sixth International Conference on Language Resources and Evaluation*, 2008.
[16] C. Huor, T. Rithy, R. Hemy, V. Navy, C. Chanthirith, and C. Tola, "Word bigram vs orthographic syllable bigram in khmer word," in *PAN Localization Team, Cambodia*, 2007.
[17] N. Bi and N. Taing, "Khmer word segmentation based on bi-directional maximal matching for plaintext and microsoft word document," in *Proc. APSIPA ASC*, 2014.
[18] V. Chea, Y. Thu, C. Ding, M. Utiyama, A. Finch, and E. Sumita, "Khmer word segmentation using conditional random fields," in *Proc. Khmer Natural Language Processing (KNLP)*, 2015.
[19] Nat. Chuon, "Khmer electronic dictionary (grammar part)," in *Buddhist Institute Dictionary*, 1967.
[20] An. Stolcke, "Srilm: an extensible language modeling toolkit," in *Proc. International conference of spoken language processing (ICSLP)*, 2002.
[21] D. Povey and et al., "The Kaldi speech recognition toolkit," in *Proc. IEEE-ASRU*, 2011.
[22] P.R. Dixon, C. Hori, and H. Kashioka, "Development of the SprinTra WFST speech decoder," *NICT Research Journal*, pp. 15–20, 2012.
[23] H. Naing, A. Hlaing, W. Pa, X. Hu, Y. Thu, C. Hori, and H. Kawai, "A myanmar large vocabulary continuous speech recognition system," in *Proc. Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2015.