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

Investigating Radical-based End-to-End Speech Recognition Systems for Chinese Dialects and Japanese

Sheng Li[1]*, *Xugang Lu*[1], *Chenchen Ding*[1]*, *Peng Shen*[1], *Tatsuya Kawahara*[1,2]*, *Hisashi Kawai*[1]

[1]National Institute of Information and Communications Technology, Kyoto, Japan
[2]Kyoto University, Kyoto, Japan

sheng.li@nict.go.jp

Abstract

Training automatic speech recognition (ASR) systems for East Asian languages (e.g., Chinese and Japanese) is tough work because of the characters existing in the writing systems of these languages. Traditionally, we first need to get the pronunciation of these characters by morphological analysis. The end-to-end (E2E) model allows for directly using characters or words as the modeling unit. However, since different groups of people (e.g., residents in Chinese mainland, Hong Kong, Taiwan, and Japan) adopts different writing forms for a character, this also leads to a large increase in the number of vocabulary, especially when building ASR systems across languages or dialects. In this paper, we propose a new E2E ASR modeling method by decomposing the characters into a set of radicals. Our experiments demonstrate that it is possible to effectively reduce the vocabulary size by sharing the basic radicals across different dialect of Chinese. Moreover, we also demonstrate this method could also be used to construct a Japanese E2E ASR system. The system modeled with radicals and kana achieved similar performance compared to state-of-the-art E2E system built with word-piece units.

Index Terms: Speech recognition, acoustic model, end-to-end model, radicals, transformer

1. Introduction

Conventional GMM-HMM [1] and DNN-HMM [2] based automatic speech recognition (ASR) systems require independently optimized components: acoustic model, lexicon and language model. The end-to-end (E2E) model integrates these components into a single neural network. It simplifies ASR system construction, solves the sequence labeling problem between variable-length speech frame inputs and label outputs (phone, character, syllable, word, etc.) and has achieved promising results on ASR tasks. Various types of E2E model have been studied in recent years: connectionist temporal classification (CTC) [3, 4], attention-based encoder-decoder (Attention) E2E models [5, 6], E2E lattice-free maximum mutual information (LFMMI) [7], and E2E models jointly trained with CTC and attention-based objectives (CTC/Attention) [8, 9, 10, 11]. Recently, the transformer [12] has been applied to E2E speech recognition systems [13, 14, 15, 16] and has achieved promising results.

However, building multi-dialect/multi-lingual End-to-End ASR systems for East Asian languages directly using characters is still an intricate problem due to a large number of existing character categories. There are several character sets coexisting in Chinese, Japanese and Korean languages. And the total number of the characters is more than 20,000. As an alternative, most conventional approaches only recognize about 2,000 to 4,000 commonly used characters, and inevitably have problems of OOV.

In this paper, we propose a novel radical-based representation for Chinese characters. It is well-known that all Chinese characters are composed of basic structural components, called radicals. Thus, these large characters can be decomposed into a compact set of basic radicals. The manner of treating a Chinese character as a composition of radicals rather than a single character class largely reduces the size of vocabulary. Therefore, it is an intuitive way to decompose Chinese characters into radicals and describe their spatial structures for E2E acoustic modeling. Compared with traditional character-based methods, the main contributions of this study are summarized as follows: We describe how to decomposite Chinese characters based on detailed analysis of Chinese radicals and structures. The size of the radical vocabulary is largely reduced compared with the character vocabulary. We also effectively build multi-dialect E2E ASR systems for accents of Chinese Mainland, Taiwan, and Hong Kong. The same method can also be applied to the Japanese language.

The remainder of this paper is organized as follows. Section 2 briefly reviews the related work and describes our proposed method. Section 3 provides experimental evaluations with different tasks. Conclusions and future works are given in Section 4.

2. Decompositing Chinese Characters based on Structural Analysis

2.1. Catergory of Characters

As we introduced in Section 1, several character sets are coexisting in some East Asian languages:

1. Traditional Chinese characters ("Zhengti" or "Fanti") are still used primarily in Taiwan, Macau, Hong Kong, and many overseas communities. Moreover, they also remain in use in mainland China for artistic, scholarly and advertising purposes.
2. Simplified Chinese characters ("Jianti") become the written form of the Chinese mainland and Singapore.
3. "Kanji", which remain a key component used alongside the Japanese syllabic scripts Hiragana and Katakana in the Japanese writing system.
4. "Hanja", which are occasionally used in the writing of Korean.
5. "Chunom", which were formerly used in Vietnamese.

* Corresponding authors and main contributors.

http://dx.doi.org/10.21437/Interspeech.2019-2104

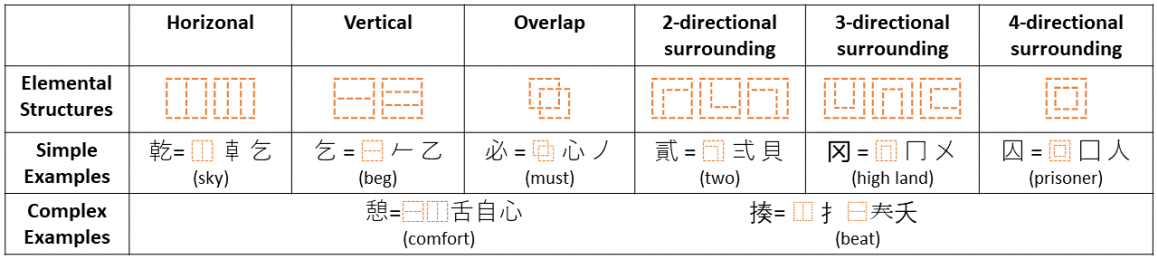

Figure 1: *The structural analysis of characters.*

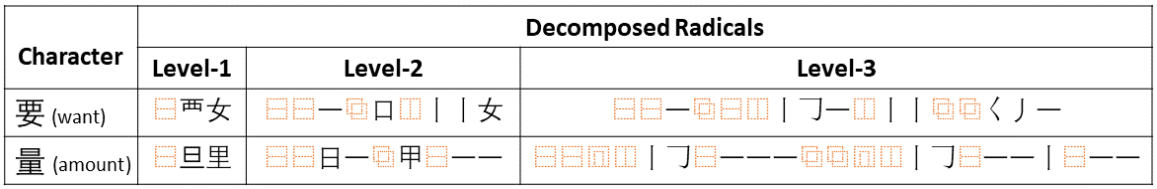

Figure 2: *Three decomposition levels.*

The above (1)-(4) are known as CJK characters [17], more than 20,000 in total. Vietnamese in (5) is sometimes also included, making the abbreviation CJKV [18], which has an even larger character size. This makes building character-based multi-dialect/multi-lingual E2E ASR systems very difficult due to a large number of existing character categories (more than 20,000), not including recently created characters from the Internet.

2.2. Related Works

In the past few decades, lots of efforts have been made for studying radical-based Chinese character decomposition.

The most in-depth research about this topic is from the input method area. The Wubi ("five radicals") input method[1] is based on the structure of characters rather than their pronunciation. In this input method, each character has a unique representation with at most four radicals. Some characters can even be written with fewer radicals. Unlike with traditional phonetic input methods, one does not have to spend time selecting the desired character from a list of homophonic possibilities. There are reports of experienced typists reaching 160 characters per minute with Wubi.

In the OCR field, [19, 20] introduced methods to first detect radicals and then to recognize a Chinese character. Recently, [21] also tried to detect position-dependent radicals using a deep residual network.

2.3. Radical-based Chinese Character Decomposition

As shown in Figure 1, we summarize twelve basic structures for characters. They are listed as follows:

- Two "**Horizonal**" structures
- Two "**Vertical**" structures
- One "**Overlap**" structure
- Three "**2-directional surrounding**" structures
- Three "**3-directional surrounding**" structures
- One "**4-directional surrounding**" structure

Following these structures, we can decompose the characters into particles. We name these particles using a professional term "Radical". The decomposition rules are all from the existing open-source project[2]. These rules are written according to Backus-Naur Form (BNF) [22, 23] style as shown in following equations.

$$\textbf{CHAR} ::= \textbf{STRUCTURE}\ \ \textbf{RAD}\ \ \textbf{RAD}\ \ [\textbf{RAD}]$$

$$\textbf{RAD} ::= \textbf{STRUCTURE}\ \ \textbf{RAD}\ \ \textbf{RAD}\ \ [\textbf{RAD}]$$

where the **CHAR** is the character to be decomposed, and the **STRUCTURE** represents one of the twelve structural-marks corresponding elemental structure in Figure 1. Two and one optional **RAD**s follow the structural-mark (**STRUCTURE**). The **RAD** is the decomposed particle which can also be iteratively decomposed into structural-mark and sub-particles.

We investigate three different levels of complexities for decomposition as shown in Figure 2.

- Level-1 (L1): We use the most direct way to decompose characters, mostly one structural-mark following two radicals as shown in Figure 1. If there is no direct decomposition, we find the decomposition method from L2, or even L3.
- Level-2 (L2): The decomposition stops until all of the radicals are the predefined components of characters (also known as "ideographs"). The total number of the ideographs is 249, which is a subset from CNS1163 character set defined by Taiwan Government[3]. If there is no decomposition found in this level, we find the decomposition method from L3.
- Level-3 (L3): It is the most aggressive decomposition. The decomposition stops until all of the radicals are the

[1] http://wubi.free.fr

[2] https://github.com/amake/cjk-decomp

[3] http://www.cns11643.gov.tw

Table 1: *Number of charaters and radicals from different decomposition levels caculated from corpora in Section 3.2 and Section 3.3*

Charsets	Simplified Chinese	Traditional Chinese	Simplified +Traditional	Kanji Japanese
#Char	2569	2651	3583	2743
↓	↓	↓	↓	↓
#L1	1072	1276	1474	1000
↓	↓	↓	↓	↓
#L2	249	242	249	242
↓	↓	↓	↓	↓
#L3	28	28	28	28

elementary radicals (also known as "strokes"). The total number of the strokes is 28, which is a simplified set from CNS1163.

Table 1 shows the number of characters and decomposed radicals from different levels calculated from the Chinese dialectal corpora in Section 3.2 and the Japanese corpus in Section 3.3. The basic 28 strokes are shared by both the Chinese (simplified and traditional) and Japanese Kanji. The basic ideographs are almost the same among these three charsets. For L1 decomposition, the numbers of radicals are about half of their corresponding character numbers.

3. Experimental Evaluations

In this section, we evaluate a set of models built with our proposed method.

3.1. Transformer-based E2E ASR systems

The ASR-Transformer maps an input sequence, that is, the log-Mel filterbank feature, to a sequence of intermediate representations by the encoder. The decoder generates an output sequence of symbols (phones, syllables, sub-words, or words) given the intermediate representations. The biggest difference between the ASR-Transformer and commonly used E2E models [5, 6] is that the ASR-Transformer completely relies on attention and feedforward components [12].

We used the implementation of the Transformer-based neural machine translation (NMT-Transformer) [12] in tensor2tensor [4] for all our experiments. The training and testing settings listed in Table 2 were similar to those in [16].

3.2. Chinese Speech Recognition Task

Traditionally, there are many regional dialects for the Chinese language. John Hopkins Summer Workshop[5] had a special report on this topic. This report [24] analyzed the temporal, frequential and prosodic differences of dialects and their influences to the LVCSR performance.

In this paper, we focus on three major groups of dialects as shown in Table 3. Together with Beijing dialect (also known as Mandarin, MA) which uses the simplified character, the others are Taiwan dialect (TW) and Hong Kong dialect (HK), which use traditional characters.

We select around two hours from these dialect speech data as the test set, so the amounts of these dialects are averaged. The other data are used as the training set. We use the Beijing dialect data (MA) for fine-tuning our proposed decomposition method. We empirically select best modeling units and test our proposed method for simplified Chinese using test set of Beijing dialect (MA). These datasets are clean speech recorded on desktop environment. They are on travel topic.

We used 120-dim filterbank features (40-dim static $+\Delta$ $+\Delta\Delta$), which were mean and variance normalized per speaker, and four frames were spliced (four left, one current and zero right). Speed-perturbation [25] was not used to save training time. To initalize these models, we use a Mandarin transformer-based model (eight head-attention, six encoder-blocks and six decoder-blocks with 512 nodes) trained from 178 hours of speech data selected from AIShell dataset [26] with the CER of 9.0% on its own evaluation set.

3.2.1. Selecting Reliable Decomposition Strategy

We use the speech data of Beijing dialect for selecting the most appropriate character decomposition strategy. As we introduced in Section 2, we conduct the proposed decomposition in three different levels of complexities. Then, we train a set of radical-based E2E models using the training set of Beijing dialect. These E2E models are using the radical-based labels decomposed from characters from Level-1 (L1) to Level-3 (L3). For each decomposition level, we train two models with and without the structure marks introduced in Section 2.3. We will choose the decomposition method with the best performance on the test set of the Beijing dialect (MA). Moreover, we also checked whether the structure marks are necessary.

The results in Table 4 show the Level-1 (L1) decomposition is most reliable for constructing E2E ASR systems. We reason that the L1 decomposition is moderate and it preserves the most substantial information correlated to the character. According

Table 2: *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 3: *Desktop Recording from Three Dialect Areas (Traveling topic)*

Dataset	Dialect Area	Hours
Training	Beijing (MA)	34.9
	Hong Kong (HK)	39.8
	Taiwan (TW)	50.2
Testing	Beijing (MA)	0.6
	Hong Kong (HK)	0.8
	Taiwan (TW)	0.6

[4] https://github.com/tensorflow/tensor2tensor

[5] http://old-site.clsp.jhu.edu/ws04/groups/ws04casr

Table 4: *ASR performance (CER%) of radical-based E2E models with different settings on Development set (Beijing dialect)*

Levels (#Radical)	L1 (1042)	L2 (249)	L3 (28)
w/ structure mark	6.1	13.4	not working
w/o structure mark	**2.7**	6.8	71.9

Table 5: *ASR performance (CER%) of acoustic models with different settings on three-dialectal test sets ("**c**" means character-based model, "**r**" means radical(L1)-based model)*

Models		Dialectal Test Sets		
Training Sets	#units	MA	HK	TW
MA (**c**)	(#char=) 2569	**1.5**	33.7	35.5
HK (**c**)	2651	29.7	**0.9**	**0.4**
TW (**c**)	2651	30.5	**0.4**	**0.9**
MA+HK+TW (**c**)	3583	**1.5**	**0.2**	**0.1**
MA (**r**)	(#L1=) 1072	2.7	15.4	11.6
HK (**r**)	1276	27.5	4.8	4.7
TW (**r**)	1276	30.4	4.2	7.1
MA+HK+TW (**r**)	1474	**1.3**	**0.1**	**0.03**

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

to the "Modern Chinese General Character List"[27], 57.4% of the top 3000 most frequently used characters are composed with "meaning radical", which indicates the meaning of the character, and "sound radical", which indicates the pronunciation. On the other hand, the L2 and L3 decomposition methods are over-aggressive, and information related to the character (the "meaning radical" and "sound radical") could be lost during the decomposition process.

The results also show the structure marks are not helpful and it will significantly increase insertion and deletion rates. It is possible to remove them because all children in Eastern Asia are taught writing each character with only one default order (left-to-right, up-to-down and outside-to-inside). However, it could generate conflictions (different characters share the same radical sequence). By assigning the conflicting characters with hand-designed unique decompositions, CER% of the recognition can achieve reduction around 2%.

3.2.2. Multi-Dialect Chinese Speech Recognition Task

We evaluate our proposed method on the multi-dialect Chinese speech recognition task. We decompose the three dialectal Chinese datasets on Level-1 (L1). In Table 5, although single language modeling with this radical units is not so effective, the number of the decomposed radicals can be reduced by a half of the number of characters for each of these dialects. Using multi-lingual training [14], the radical-based multi-dialect model (MA+HK+TW (**r**)) can achieve almost similar performance as the character-based multi-dialect model (MA+HK+TW (**c**)), while the number of modeling units is also reduced more than half.

3.3. Japanese Speech Recognition Task

As we mentioned in Section 2, characters (Japanese call them "Kanji") has deeply embedded in the Japanese writing system since ancient times. After the industrial revolution, the words (usually character-based) about western civilization (technology, politics, philosophy, art, and laws) were first created in Japan and then spread to China. Today, the proportion of a Japanese article people from both countries can understand is more than one third. For this reason, we test our proposed method on the "Corpus of Spontaneous Japanese (CSJ)" [28]. We used approximately 577 hours of lecture recordings as the training set (**CSJ-Train**) according to [29, 30, 11, 31]. Three official evaluation sets (**CSJ-Eval01**, **CSJ-Eval02**, and **CSJ-Eval03**), each containing ten lecture recordings [31], were used to evaluate the speech recognition results. Ten lecture recordings were selected for development (**CSJ-Dev**).

Table 6: *ASR performance (CER%) of the ASR-Transformer models trained with different units*

Network	#unit	CER%			
		E01	E02	E03	Ave.
char	3178	8.2	5.9	6.6	6.9
word	98245	10.2	8.6	9.7	9.5
WPM	3000	8.4	6.1	6.3	6.9
	8000	**7.8**	**6.0**	**6.1**	**6.6**
radical(L1)+kana	1171	**8.0**	**5.8**	**6.2**	**6.7**

The results without statistical significance (two-tailed t-test at p-value < 0.05) are shown in bold fonts.

We used 72-dim filterbank features (24-dim static +Δ +$\Delta\Delta$). Other settings are the same as the series of models in Section 3.2. We trained the baseline ASR-Transformer models using CSJ-Train. For testing, we decoded the speech from test sets (CSJ-E01/02/03) and evaluated our models using the character error rate (CER%). Several modeling units were compared including words, word-piece-model (WPM)[32] and characters as shown in Table 6. We used the sentence-piece toolkit[6] as the sub-word segmenter. We used separate 3000 and 8000 sub-word vocabularies. The ASR-Transformer model trained with 1000 radicals together with 171 Kanas had similar best results compared to the model 8000 word-pieces. This means we discovered a more efficient way for training Japanese E2E model.

4. Conclusions and Future Work

The end-to-end (E2E) model allows directly using characters or words as the modeling unit. However, since different groups of people (e.g., residents in Chinese, Hong Kong, Taiwan, and Japan) adopts different writing forms for a character, this also leads to a large increase in the number of vocabulary, especially when building ASR systems across languages or dialects. In this paper, we propose a new E2E ASR modeling method by decomposing the characters into a set of radicals. Our experiments demonstrate that it is possible to effectively reduce the vocabulary size by sharing the basic radicals across different dialect of Chinese. Moreover, we also demonstrated this method could also be used to construct Japanese E2E ASR system. In the future, we will move further to model the Chinese, Korean and Japanese languages together.

5. References

[1] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*,

[6] https://github.com/google/sentencepiece

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 *CoRR abs/1706.03762*, 2017.

[13] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. IEEE-ICASSP*, 2018.

[14] S. Zhou, S. Xu, and B. Xu, "Multilingual end-to-end speech recognition with a single transformer on low-resource languages," in *CoRR abs/1806.05059*, 2018.

[15] S. Zhou, L. Dong, S. Xu, and B. Xu, "A comparison of modeling units in sequence-to-sequence speech recognition with the transformer on mandarin chinese," in *CoRR abs/1805.06239*, 2018.

[16] S.Zhou, L.Dong, S.Xu, and B.Xu, "Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin chinese," in *Proc. INTERSPEECH*, 2018.

[17] C. Leban, *Automated Orthographic Systems for East Asian Languages (Chinese, Japanese, Korean), State-of-the-art Report, Prepared for the Board of Directors*. Association for Asian Studies, 1971.

[18] K. Lunde, *CJKV Information Processing, 2nd Edition, Chinese, Japanese, Korean, and Vietnamese Computing*. O'Reilly Associates, 2009.

[19] L. Ma and C. Liu, "A new radical-based approach to online handwritten chinese character recognition," in *Proc. IEEE-International Conference on Pattern Recognition (CVPR)*, 2008.

[20] A. Wang and K. Fan, "Optical recognition of handwritten Chinese characters by hierarchical radical matching method," *Pattern Recognition*, pp. 15–35, 2001.

[21] T. Wang, F. Yin, and C. Liu, "Radical-based chinese character recognition via multi-labeled learning of deep residual networks," in *Proc. International Conference on Document Analysis and Recognition (ICDAR)*, 2017.

[22] J. Backus, "The syntax and semantics of the proposed international algebraic language of the zurich acm-gamm conference," in *Proc. International Conference on Information Processing, UNESCO*, 1959, pp. 125–132.

[23] P. Naur, "A course on algol programming," in *Note 1, Retrieved 26 March 2015*, 1961, p. 5.

[24] R. Sproat, T. F. Zheng, L. Gu, D. Jurafsky, I. Shanfran, J. Li, Y. Zheng, H. Zhou, Y. Su, S. Tsakalidis, P. Bramsen, and D. Kirsch, "Dialectal chinese speech recognition: Final technical report," in *Summer Workshop at CLSP/JHU Technical Report*, 2004.

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

[26] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AIShell-1: An open-source Mandarin speech corpus and a speech recognition baseline," in *Proc. Oriental COCOSDA*, 2017.

[27] *Modern Chinese General Character List*. National Language and Literature Working Committee and Publication Administration of the People's Republic of China, 1988.

[28] K. Maekawa, "Corpus of spontaneous japanese: Its design and evaluation," in *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[29] T. Moriya, T. Shinozaki, and S. Watanabe, "Kaldi recipe for Japanese spontaneous speech recognition and its evaluation," in *Autumn Meeting of ASJ*, no. 3-Q-7, 2015.

[30] N. Kanda, X. Lu, and H. Kawai, "Maximum a posteriori based decoding for CTC acoustic models," in *Proc. INTERSPEECH*, 2016, pp. 1868–1872.

[31] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui, "Benchmark test for speech recognition using the corpus of spontaneous japanese," in *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[32] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," in *CoRR abs/1804.10959*, 2018.