Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2622–2628
Marseille, 11–16 May 2020

Speech Corpus of Ainu Folklore and End-to-end Speech Recognition for Ainu Language

Kohei Matsuura, Sei Ueno, Masato Mimura, Shinsuke Sakai, Tatsuya Kawahara
Graduate School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
{matsuura, ueno, mimura, sakai, kawahara}@sap.ist.i.kyoto-u.ac.jp

Abstract

Ainu is an unwritten language that has been spoken by Ainu people who are one of the ethnic groups in Japan. It is recognized as critically endangered by UNESCO and archiving and documentation of its language heritage is of paramount importance. Although a considerable amount of voice recordings of Ainu folklore has been produced and accumulated to save their culture, only a quite limited parts of them are transcribed so far. Thus, we started a project of automatic speech recognition (ASR) for the Ainu language in order to contribute to the development of annotated language archives. In this paper, we report speech corpus development and the structure and performance of end-to-end ASR for Ainu. We investigated four modeling units (phone, syllable, word piece, and word) and found that the syllable-based model performed best in terms of both word and phone recognition accuracy, which were about 60% and over 85% respectively in speaker-open condition. Furthermore, word and phone accuracy of 80% and 90% has been achieved in a speaker-closed setting. We also found out that a multilingual ASR training with additional speech corpora of English and Japanese further improves the speaker-open test accuracy.

Keywords: Ainu speech corpus, low-resource language, end-to-end speech recognition

1. Introduction

Automatic speech recognition (ASR) technology has been made a dramatic progress and is currently brought to a pratical levels of performance assisted by large speech corpora and the introduction of deep learning techniques. However, this is not the case for low-resource languages which do not have large corpora like English and Japanese have. There are about 5,000 languages in the world over half of which are faced with the danger of extinction. Therefore, constructing ASR systems for these endangered languages is an important issue.

The Ainu are an indigenous people of northern Japan and Sakhakin in Russia, but their language has been fading away ever since the Meiji Restoration and Modernization. On the other hand, active efforts to preserve their culture have been initiated by the Government of Japan, and exceptionally large oral recordings have been made. Nevertheless, a majority of the recordings have not been transcribed and utilized effectively. Since transcribing them requires expertise in the Ainu language, not so many people are able to work on this task. Hence, there is a strong demand for an ASR system for the Ainu language. We started a project of Ainu ASR and this article is the first report of this project.

We have built an Ainu speech corpus based on data provided by the Ainu Museum[1] and the Nibutani Ainu Culture Museum[2]. The oral recordings in this data consist of folklore and folk songs, and we chose the former to construct the ASR model. The end-to-end method of speech recognition has been proposed recently and has achieved performance comparable to that of the conventional DNN-HMM hybrid modeling (Chiu et al., 2017; Povey et al., 2018; Han et al., 2019). End-to-end systems do not have a complex hierarchical structure and do not require expertise in target languages such as their phonology and morphology. In this study we adopt the attention mechanism (Chorowski et al., 2014; Bahdanau et al., 2016) and combine it with Connectionist Temporal Classification (CTC) (Graves et al., 2006; Graves and Jaitly, 2014). In this work, we investigate the modeling unit and utilization of corpora of other languages.

2. Overview of the Ainu Language

This section briefly overviews the background of the data collection, the Ainu language, and its writing system. After that, we describe how Ainu recordings are classified and review previous works dealing with the Ainu language.

2.1. Background

The Ainu people had total population of about 20,000 in the mid-19th century (Hardacre, 1997) and they used to live widely distributed in the area that includes Hokkaido, Sakhalin, and the Kuril Islands. The number of native speakers, however, rapidly decreased through the assimilation policy after late 19th century. At present, there are only less than 10 native speakers, and UNESCO listed their language as critically endangered in 2009 (Alexandre, 2010). In response to this situation, Ainu folklore and songs have been actively recorded since the late 20th century in efforts initiated by the Government of Japan. For example, the Ainu Museum started audio recording of Ainu folklore in 1976 with the cooperation of a few Ainu elders which resulted in the collection of speech data with the total duration of roughly 700 hours. This kind of data should be a key to the understanding of Ainu culture, but most of it is not transcribed and fully studied yet.

2.2. The Ainu Language and its Writing System

The Ainu language is an agglutinative language and has some similarities to Japanese. However, its genealogical relationship with other languages has not been clearly understood yet. Among its features such as closed syllables

[1] http://ainugo.ainu-museum.or.jp/
[2] http://www.town.biratori.hokkaido.jp/biratori/nibutani/

Table 1: Speaker-wise details of the corpus

Speaker ID	KM	UT	KT	HS	NN	KS	HY	KK	total
duration (h:m:s)	19:40:58	7:14:53	3:13:37	2:05:39	1:44:32	1:43:29	1:36:35	1:34:55	38:54:38
duration (%)	50.6	18.6	8.3	5.4	4.5	4.4	4.1	4.1	100.0
# episodes	29	26	20	8	8	11	8	7	114
# IPUs	9170	3610	2273	2089	2273	1302	1220	1109	22345

and personal verbal affixes, one important feature is that there are many compound words. For example, a word *atuykorkamuy* (means "a sea turtle") can be disassembled into *atuy* ("the sea"), *kor* ("to have"), and *kamuy* ("god").

Although the Ainu people did not traditionally have a writing system, the Ainu language is currently written following the examples in a reference book "Akor itak" (The Hokkaido Utari Association, 1994). With this writing system, it is transcribed with sixteen Roman letters {a, c, e, h, i, k, m, n, o, p, r, s, t, u, w, y}. Since each of these letters correspond to a unique pronunciation, we call them "phones" for convenience. In addition, the symbol {=} is used for connecting a verb and a personal affix and { ' } is used to represent the pharyngeal stop. For the purpose of transcribing recordings, consonant symbols {b, d, g, z} are additionally used to transcribe Japanese sounds the speakers utter. The symbols { _ , __ } are used to transcribe drops and liaisons of phones. An example is shown below.

original	mos=an __hine inkar'=an				
translation	I wake up and look				
structure	mos	=an	hine	inkar	=an
	wake up	**1sg**	**and**	**look**	**1sg**

2.3. Types of Ainu Recordings

The Ainu oral traditions are classified into three types: "*yukar*" (heroic epics), "*kamuy yukar*" (mythic epics), and "*uwepeker*" (prose tales). *Yukar* and *kamuy yukar* are recited in the rhythm while *uwepeker* is not. In this study we focus on the the prose tales as the first step.

2.4. Previous Work

There have so far been a few studies dealing with the Ainu language. Senuma and Aizawa (2018) built a dependency tree bank in the scheme of Universal Dependencies[3]. Nowakowski et al. (2019) developed tools for part-of-speech (POS) tagging and word segmentation. Ainu speech recognition was tried by Anastasopoulos and Chiang (2018) with 2.5 hours of Ainu folklore data even though the Ainu language was not their main target. Their phone error rare was about 40% which is not an accuracy level for practical use yet.

It appears that there has not been a substantial Ainu speech recognition study yet that utilizes corpora of a reasonable size. Therefore, our first step was to build a speech corpus for ASR based on the data sets provided by the Ainu Museum and the Nibutani Ainu Culture Museum.

[3]https://universaldependencies.org/

Table 2: Text excerpted from the prose tale '*The Boy Who Became Porosir God*' spoken by KM.

original	*English translation*
Samormosir mosir	In neighboring country
noski ta	at the middle (of it),
a=kor hapo i=resu hine	being raised by my mother,
oka=an pe ne _hike	I was leading my life.
kunne hene tokap _hene	Night and day, all day long,
yam patek i=pareoyki	I was fed with chestnut
yam patek a=e kusu	and all I ate was chestnut,
somo hetuku=an pe ne kunak	so, that I would not grow up
a=ramu a korka	was my thought.

3. Ainu Speech Corpus

In this section we explain the content of the data sets and how we modified it for our ASR corpus.

3.1. Numbers of Speakers and Episodes

The corpus we have prepared for ASR in this study is composed of text and speech. Table 1 shows the number of episodes and the total speech duration for each speaker. Among the total of eight speakers, the data of the speakers KM and UT is from the Ainu Museum, and the rest is from Nibutani Ainu Culture Museum. All speakers are female. The length of the recording for a speaker varies depending on the circumstances at the recording times. A sample text and its English translation are shown in Table 2.

3.2. Data Annotation

For efficient training of ASR model, we have made some modifications to the provided data. First, from the transcripts explained in Section 2.1, the symbols {_ , __ , '} have been removed as seen in the example below.

original	uymam'=an wa isam=an __hi okake ta
modified	uymam=an wa isam=an hi okake ta

Though the equal symbol ('=') does not represent a sound, we keep it because it is used in almost all of the Ainu documents and provides grammatical information.

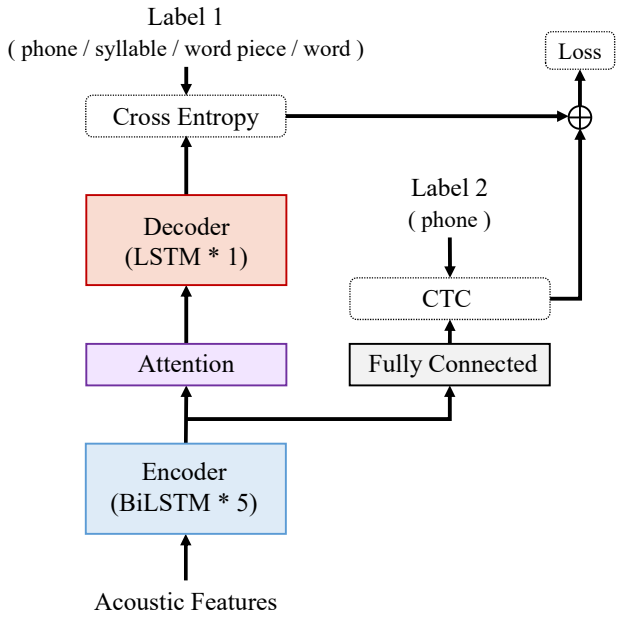

Figure 1: The attention model with CTC auxiliary task.

To train an ASR system, the speech data needs to be segmented into a set of manageable chunks. For the ease of automatic processing, we chose to segment speech into *interpausal units* (IPUs) (Koiso et al., 1998)which is a stretch of speech bounded by pauses. The number of IPUs for each speaker is shown in Table 1.

4. End-to-end Speech Recognition

In this section, the two approaches to end-to-end speech recognition that we adopt in this work are summarized. Then, we introduce four modeling units we explained, i.e., phone, syllable, word piece, and word. We also discuss multilingual training that we adopt for tackling the low resource problem.

4.1. End-to-end Modeling

End-to-end models have an architecture much simpler than that of conventional DNN-HMM hybrid models. Since they predict character or word symbols directly from acoustic features, pronunciation dictionaries and language modeling are not required explicitly. In this paper, we utilize two kinds of end-to-end models, namely, Connectionist Temporal Classification (CTC) and the attention-based encoder-decoder model.

CTC augments the output symbol set with the "blank" symbol 'ϕ'. It outputs symbols by contracting frame-wise outputs from recurrent neural networks (RNNs). This is done by first collapsed repeating symbols and then removing all blank symbols as in the following example:

$$\text{aab}\phi\text{bbccc} \rightarrow \text{abbc}$$

The probability of an output sequence $\boldsymbol{L}$ for an input acoustic feature sequence $\boldsymbol{X}$ ($|\boldsymbol{L}| < |\boldsymbol{X}|$) is defined as follows.

$$p(\boldsymbol{L}|\boldsymbol{X}) = \sum_{\substack{\boldsymbol{\Pi} \in \mathcal{B}^{-1}(\boldsymbol{L}) \\ |\boldsymbol{\Pi}| = |\boldsymbol{X}|}} p(\boldsymbol{\Pi}|\boldsymbol{X}) \tag{1}$$

$\mathcal{B}$ is a function to contract the outputs of RNNs, so $\mathcal{B}^{-1}(\boldsymbol{L})$ means the set of symbol sequences which is reduced to $\boldsymbol{L}$. The model is trained to maximize (1).

The attention-based encoder-decoder model is another method for mapping between two sequences with different lengths. It has two RNNs called the "encoder" and the "decoder". In naive encoder-decoder model, the encoder converts the input sequence into a single context vector which is the last hidden state of the encoder RNN from which the decoder infers output symbols. In an attention-based model, the context vector $\mathbf{c}_l$ at l-th decoding step is the sum of the product of all encoder outputs $\boldsymbol{h}_1, ..., \boldsymbol{h}_{\mathrm{T}}$ and the l-th attention weight $\alpha_{1,l}, ..., \alpha_{\mathrm{T},l}$ as shown in (2). Here, T is the length of the encoder output.

$$\boldsymbol{c}_l = \sum_{t=1}^{\mathrm{T}} \alpha_{t,l} \boldsymbol{h}_t \tag{2}$$

The attention weights $\boldsymbol{\alpha}_{1,l}, ..., \boldsymbol{\alpha}_{\mathrm{T},l}$ indicates the relative importances of the encoder output frames for the l-th decoding step and the model parameters to generate these weights are determined in an end-to-end training.

In our model, the attention-based model and the CTC share the encoder and are optimized simultaneously as shown in Figure 1.(Kim et al., 2016) Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) is used for RNNs in the encoder and the decoder.

4.2. Modeling Units

In the conventional DNN-HMM hybrid modeling, the acoustic model outputs probabilities triphone states from each acoustic feature which is converted into the most likely word sequence. An end-to-end model, on the other hand, has some degree of freedom in the modeling unit other than phones, and there are some studies that use characters or words as a unit (Chan et al., 2015; Li et al., 2018). A word unit based end-to-end model can take long context into consideration at the inference time, but it has the data sparsity problem due to its large vocabulary size. Though phone unit based model does not have such a problem, it cannot grasp so long context. It depends on the size of available corpora to decide which to adopt. In addition to these both models, a word piece unit, which is defined by automatically dividing a word into frequent parts, has been proposed (Schuster and Nakajima, 2012; Lüscher et al., 2019), and its vocabulary size can be determined almost freely.

In this paper, we investigate the modeling unit for the end-to-end Ainu speech recognition since the optimal unit for this size of corpus is not obvious. (Irie et al., 2019) It is presupposed that all units can be converted into word units automatically. The candidates are phone, syllable, word piece (WP), and word. Examples of them are shown in Table 3 and the details of each unit are described below.

Table 3: Examples of four modeling units.

original	a=saha i=kokopan wa
phone	a = s a h a ⟨wb⟩ i = k o k o p a n ⟨wb⟩ w a
syllable	a = sa ha ⟨wb⟩ i = ko pan ⟨wb⟩ wa
WP	⟨wb⟩a = saha ⟨wb⟩i = ko p an ⟨wb⟩wa
word	a = saha i = ⟨unk⟩ wa
translation	my elder sister told me not to do so

4.2.1. Phone

As mentioned in Section 2.1, we regard the Roman letters as phones. '=' and the special symbol '⟨wb⟩', which means a word boundary, are added to make it possible to convert the output into a sequence of words like the 'original' in Table 3.

4.2.2. Syllable

A syllable of the Ainu language takes the form of either V, CV, VC, or CVC, where 'C' and 'V' mean consonant and vowel, respectively. The phones {a, e, i, o, u} are vowels and the rest of the Roman letters in Section 2.2 are consonants. In this work, every word is divided into syllables by the following procedure.

1. A word with a single letter is unchanged.
2. Two consecutive Cs and Vs are given a syllable boundary between them.

$$\mathrm{R}^*\{\mathrm{CC},\ \mathrm{VV}\}\mathrm{R}^* \rightarrow \mathrm{R}^*\{\mathrm{C\text{-}C},\ \mathrm{V\text{-}V}\}\mathrm{R}^* \quad (\mathrm{R} := \{\mathrm{C},\ \mathrm{V}\})$$

3. Put a syllable boundary after the segment-initial V if it is following by at least two phones.

$$\mathrm{VCR}^+ \rightarrow \mathrm{V\text{-}CR}^+$$

4. Put a syllable boundary after CV repeatedly from left to right until only CV or CVC is left.

$$(\mathrm{CV})^*\{\mathrm{CV},\ \mathrm{CVC}\} \rightarrow (\mathrm{CV\text{-}})^*\{\mathrm{CV},\ \mathrm{CVC}\}$$

In addition, '=' and '⟨wb⟩' are added as explained in Section 4.2.1. through the model training process.
This procedure does not always generate a morphologically relevant syllable segmentation. For example, a word *iser-makus* (meaning "(for a god) to protect from behind") is divided as *i-ser-ma-kus*, but the right syllabification is *i-ser-mak-us*.

4.2.3. Word Piece

The byte pair encoding (BPE) (Sennrich et al., 2015) and the unigram language modeling (Kudo, 2018) are alternative methods for dividing a word into word pieces. The former repeatedly replaces the most common character pair with a new single symbol until the vocabulary becomes the intended size. The latter decides the segmentation to maximize the likelihood of occurrence of the sequence. We adopt the latter and use the open-source software SentencePiece[4] (Kudo and Richardson, 2018). With this tool, '⟨wb⟩' and other units are often merged to constitute a single piece as seen in Table 3.

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

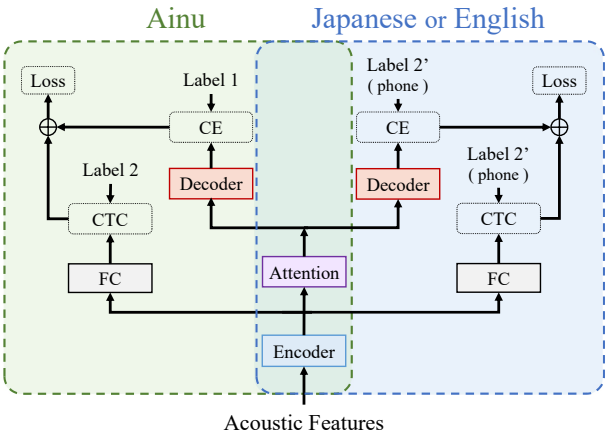

Figure 2: The architecture of the multilingual learning with two corpora. 'FC' and 'CE' means 'fully connected' and 'cross-entropy' respectively.

4.2.4. Word

The original text can be segmented into words separated by spaces. To make the vocabulary smaller for the ease of training, '=' is treated as a word and infrequent words are replaced with a special label '⟨unk⟩'. As seen in Table 3, '*a=saha*' is dealt with as three words ('*a*', '=', '*saha*') and the word '*kokopan*' is replaced with '⟨unk⟩'.

4.3. Multilingual Training

When an enough amount of data is not available for the target languages, the ASR model training can be enhanced by taking advantage of data from other languages (Toshniwal et al., 2018; Cho et al., 2018). There are some similarities between Ainu and Japanese language (Tamura, 2013). For instance, both have almost the same set of vowels and do not have consonant clusters (like '*str*' of 'strike' in English). Hence, the multilingual training with a Japanese corpus is expected to be effective. In addition, an English corpus is used for the purpose of comparison. The corpora used are the JNAS corpus (Itou et al., 1999) (in Japanese) and the WSJ corpus (Paul and Baker, 1992) (in English). JNAS comprises roughly 80 hours from 320 speakers, and WSJ has about 70 hours of speech from 280 speakers.

In the multilingual training, the encoder and the attention module are shared among the Ainu ASR model and the models for other languages, and they are trained using data for all languages. Figure 2 shows the architecture for the multilingual learning with two corpora. When the input acoustic features are from the Ainu ASR corpus, they go through the shared encoder and attention module and are delivered into the decoder on the left side in Figure 2 as a context vector. In this case, the right-side decoder is not trained.

Table 4: ASR performance for each speaker and modeling unit. The lowest error rates for each unit are highlighted.

		units	KM	UT	KT	HS	NN	KS	HY	KK	average
speaker-closed	WER (%)	phone	22.2	28.5	24.2	28.6	27.2	30.6	40.4	36.1	27.9
		syllable	**13.2**	**18.4**	**19.6**	29.4	26.7	26.7	38.9	**29.0**	**21.7**
		WP	14.4	20.0	21.6	**25.0**	**27.1**	**23.2**	**37.8**	42.5	22.3
		word	14.7	19.6	21.3	32.9	**27.1**	24.6	40.7	31.2	23.1
	PER (%)	phone	10.7	16.3	7.9	**5.6**	**7.4**	13.6	10.1	14.8	11.1
		syllable	**3.2**	**6.9**	**4.4**	7.7	7.9	9.5	**9.4**	**10.7**	**6.3**
		WP	4.7	8.0	5.2	6.7	8.4	**6.8**	10.4	12.6	7.1
		word	11.2	12.9	12.6	24.0	17.1	15.4	27.0	20.1	15.9
speaker-open	WER (%)	phone	-	-	38.8	40.5	41.9	53.1	35.9	54.7	43.4
		syllable	-	-	**33.4**	37.8	**37.3**	**47.2**	32.0	**48.6**	**38.6**
		WP	-	-	58.4	**37.2**	38.6	47.9	32.6	48.8	45.7
		word	-	-	34.0	49.0	39.4	48.9	**31.5**	84.3	46.6
	PER (%)	phone	-	-	14.9	13.9	15.9	21.4	11.2	27.0	17.1
		syllable	-	-	**10.7**	**12.6**	**13.5**	**16.5**	**10.3**	**22.0**	**13.8**
		WP	-	-	41.5	14.1	15.9	19.3	11.5	23.6	23.6
		word	-	-	24.6	39.9	29.6	33.1	20.4	67.0	34.8

5. Experimental Evaluation

In this section the setting and results of ASR experiments are described and the results are discussed.

5.1. Data Setup

The ASR experiments were performed in speaker-open condition as well as speaker-closed condition.

In the speaker-closed condition, two episodes were set aside from each speaker as development and test sets. Thereafter, the total sizes of the development and test sets turns out to be 1585 IPUs spanning 2 hours 23 minutes and 1841 IPUs spanning 2 hours and 48 minutes respectively. The ASR model is trained with the rest data. In the speaker-open condition, all the data except for the test speaker's were used for training As it would be difficult to train the model if all of the data of speaker KM or UT were removed, experiments using their speaker-open conditions were not conducted.

5.2. Experimental Setting

The input acoustic features were 120-dimensional vectors made by frame stacking (Tian et al., 2017) three 40-dimensional log-mel filter banks features at contiguous time frames. The window length and the frame shift were set to be 25ms and 10ms. The encoder was composed of five BiLSTM layers and the attention-based decoder had a single layer of LSTM. Each LSTM had 320 cells and their weights were randomly initialized using a uniform distribution He et al. (2015) with biases of zero. The fully connected layers were initialized following $\mathcal{U}(-0.1, 0.1)$. The weight decay (Krogh and Hertz, 1992) whose rate was 10^{-5} and the dropout (Srivastava et al., 2014) following $\mathcal{B}e(0.2)$ were used to alleviate overfitting. The parameters were optimized with Adam (Kingma and Ba, 2014). The learning rate was 10^{-3} at first and was multiplied by 10^{-1} at the beginning of 31st and 36th epoch (You et al., 2019). The mini-batch size was 30 and the utterances (IPUs) were sorted in an ascending order of length. To stabilize the training, we removed utterances longer than 12 seconds.

The loss function of the model was a linear sum of the loss from CTC and the attention-based decoder,

$$\mathcal{L}_{\text{all}} = \lambda \mathcal{L}_{\text{attn}} + (1 - \lambda)\mathcal{L}_{\text{CTC}}, \tag{3}$$

where λ was set to be 0.5. Through all experiments, the phone labels are used to train the auxiliary CTC task because it is reported that the hierarchical architecture, using few and general labels in the auxiliary task, improves the performance (Sanabria and Metze, 2018).

Strictly speaking, the number of each modeling units depends on the training set, but there are roughly 25-phone, 500-syllable, and 5,000-word units including special symbols that represent the start and end of a sentence. The words occurring less than twice were replaced with '⟨unk⟩'. The vocabulary size for word piece modeling was set to be 500. These settings were based on the results of preliminary experiments with the development set.

For the multilingual training, we made three training scripts by concatenating the script of Ainu and other languages (JNAS, WSJ, JNAS and WSJ). The model was trained by these scripts until 30th epoch. From 31$^{\text{st}}$ and 40th epoch, the model was fine-turned by the Ainu script. Phone units are used for JNAS and WSJ throughout the experiments.

5.3. Results

Table 4 shows the phone error rates (PERs) and word error rates (WERs) for the speaker-closed and speaker-open settings. The 'average' is weighted by the numbers of tokens in the ground truth transcriptions for speaker-wise evaluation sets.

The word recognition accuracy reached about 80% in the speaker-closed setting. In the speaker-open setting it was 60% on average and varied greatly from speaker to speaker

Table 5: Results of multilingual training.

speaker-		closed	open
WER (%)	Ainu	21.7	38.6
	+ JNAS	**21.1**	34.8
	+ WSJ	21.3	35.8
	+ both	21.4	**34.7**
PER (%)	Ainu	6.3	13.8
	+ JNAS	**6.0**	11.7
	+ WSJ	**6.0**	12.1
	+ both	**6.0**	**11.2**

(from 50% to 70%). The best phone accuracies in the speaker-closed and speaker-open settings were about 94% and 86%. Regardless of the settings, the syllable-based modeling yielded the best WER and PER. This suggests that syllables provide reasonable coverage and constraints for the Ainu language in a corpus of this size.

The PERs of the word unit model were larger than those of other units. This is because the word model often outputs the '⟨unk⟩' symbols while other unit models are able to output symbols similar in sound as below.

ground-truth	i okake un a unuhu a onaha
syllable model	piokake un a unuhu a onaha
word model	⟨unk⟩ un a unuhu a onaha

In this example, the PER of the syllable model is 5% and that of the word model is 30% even though the WERs are the same. (The output of the syllable model is rewritten into words using the '⟨wb⟩' symbol.)

WERs are generally much larger than PERs and it is further aggravated with the Ainu language. This is because, as mentioned in Section 2.1, the Ainu language has a lot of compound words and the model may be confused about whether the output is multiple words or a single compound word. The actual outputs frequently contain errors as below. The WER of this example is 57% though the PER is zero.

ground-truth	nen poka apkas an mak an kusu
output	nenpoka apkas an makan kusu

The results of multilingual training in which the modeling unit is syllables are presented in Table 5. All error rates are the weighted averages of all evaluated speakers. Here, '+ both' represents the result of training with both JNAS and WSJ corpora. The multilingual training is effective in the speaker-open setting, providing a relative WER improvement of 10%. The JNAS corpus was more helpful than the WSJ corpus because of the similarities between Ainu and Japanese language.

6. Summary

In this study, we first developed a speech corpus for Ainu ASR and then, using the end-to-end model with CTC and the attention mechanism, compared four modeling units: phones, syllables, word pieces, and words. The best performance was obtained with the syllable unit, with which WERs in the speaker-closed and speaker-open settings were respectively about 20% and 40% while PERs were about 6% and 14%. Multilingual training using the JNAS improved the performance in the speaker-open setting. Future tasks include reducing the between-speaker performance differences by using speaker adaptation techniques.

7. Acknowledgement

The data sets used in this study are provided by the Ainu Museum and Nibutani Ainu Culture Museum. The authors would like to thank Prof. Osami Okuda of Sapporo Gakuin University for his useful advices on the Ainu language.

8. References

Alexandre, M. C. N. (2010). *Atlas of the World's Languages in Danger, 3rd edn*. Paris, UNESCO Publishing.

Anastasopoulos, A. and Chiang, D. (2018). Tied multitask learning for neural speech translation. In *Proc. NAACL HLT*, volume 1, pages 82–91.

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., and Bengio, Y. (2016). End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4945–4949, March.

Chan, W., Jaitly, N., Le, Q. V., and Vinyals, O. (2015). Listen, attend and spell. *CoRR*, abs/1508.01211.

Chiu, C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gonina, K., Jaitly, N., Li, B., Chorowski, J., and Bacchiani, M. (2017). State-of-the-art speech recognition with sequence-to-sequence models. *CoRR*, abs/1712.01769.

Cho, J., Baskar, M. K., Li, R., Wiesner, M., Mallidi, S. H. R., Yalta, N., Karafiát, M., Watanabe, S., and Hori, T. (2018). Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling. *CoRR*, abs/1810.03459.

Chorowski, J., Bahdanau, D., Cho, K., and Bengio, Y. (2014). End-to-end continuous speech recognition using attention-based recurrent nn: First results. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

Graves, A. and Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In *ICML*.

Graves, A., Fernández, S., Gomez, F. J., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*.

Han, K. J., Prieto, R., Wu, K., and Ma, T. (2019). State-of-the-art speech recognition using multi-stream self-attention with dilated 1d convolutions. In *arXiv preprint arXiv:1910.00716*.

Hardacre, H. (1997). *New directions in the study of Meiji Japan*. Brill, Leiden New York.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, abs/1502.01852.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.

Irie, K., Prabhavalkar, R., Kannan, A., Bruguier, A., Rybach, D., and Nguyen, P. (2019). Model unit exploration for sequence-to-sequence speech recognition. *CoRR*, abs/1902.01955.

Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Kobayashi, T., Shikano, K., and Itahashi, S. (1999). Jnas: Japanese speech corpus for large vocabulary continuous speech recognition research. *Journal of the Acoustical Society of Japan (E)*, 20(3):199–206.

Kim, S., Hori, T., and Watanabe, S. (2016). Joint ctc-attention based end-to-end speech recognition using multi-task learning. *CoRR*, abs/1609.06773.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., and Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. *Language and Speech*, 41(3-4):295–321. PMID: 10746360.

Krogh, A. and Hertz, J. A. (1992). A simple weight decay can improve generalization. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 4*, pages 950–957. Morgan Kaufmann.

Kudo, T. and Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226.

Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. *CoRR*, abs/1804.10959.

Li, J., Ye, G., Das, A., Zhao, R., and Gong, Y. (2018). Advancing acoustic-to-word CTC model. *CoRR*, abs/1803.05566.

Lüscher, C., Beck, E., Irie, K., Kitza, M., Michel, W., Zeyer, A., Schlüter, R., and Ney, H. (2019). RWTH ASR systems for librispeech: Hybrid vs attention - w/o data augmentation. *CoRR*, abs/1905.03072.

Nowakowski, K., Ptaszynski, M., Masui, F., and Momouci, Y. (2019). Applying support vector machines to pos tagging of the ainu language. *Proc. Computational Methods for Endangered Languages*, 2(4).

Paul, D. B. and Baker, J. M. (1992). The design for the wall street journal-based CSR corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., and Khudanpur, S. (2018). Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Proc. Interspeech 2018*, pages 3743–3747.

Sanabria, R. and Metze, F. (2018). Hierarchical multi task learning with CTC. *CoRR*, abs/1807.07104.

Schuster, M. and Nakajima, K. (2012). Japanese and korean voice search. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Senuma, H. and Aizawa, A. (2018). Universal Dependencies for Ainu. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Tamura, S. (2013). *Ainu Go No Sekai (The World of Ainu Language)*. Yoshikawa-Kobun Kan.

The Hokkaido Utari Association. (1994). *Akor itak*. CREWS.

Tian, X., Zhang, J., Ma, Z., He, Y., and Wei, J. (2017). Frame stacking and retaining for recurrent neural network acoustic model. *CoRR*, abs/1705.05992.

Toshniwal, S., Sainath, T. N., Weiss, R. J., Li, B., Moreno, P., Weinstein, E., and Rao, K. (2018). Multilingual Speech Recognition with A Single End-To-End Model. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

You, K., Long, M., Wang, J., and Jordan, M. I. (2019). How does learning rate decay help modern neural networks? In *arXiv preprint arXiv:1908.01878*.