Evaluation of Free-form Conversation Learning Effects in a Tsugaru Dialect Speech Recognition Model

Akihiro Murakami
Hirosaki University
h24ms420@hirosaki-u.ac.jp

Masashi Imai
Hirosaki University
miyabi@hirosaki-u.ac.jp

Abstract - The Tsugaru dialect, which is a regional vernacular of Aomori Prefecture, can pose communication challenges between local residents and individuals from outside the prefecture. We are conducting research aimed at developing a bidirectional speech and text conversion system between the dialects and standard Japanese utilizing artificial intelligences. This paper presents the results of evaluating the impact of training on spontaneous Tsugaru dialect speech data to improve automatic speech recognition accuracy.

I. Introduction

The Tsugaru dialect, known for its complexity, is spoken in the Tsugaru region of Aomori Prefecture. Its unique vocabulary and pronunciation often hinder communication between local residents and individuals from outside the prefecture. This language barrier can be particularly problematic in medical settings and emergency situations, where miscommunication may lead to life-threatening consequences. Therefore, there is a pressing need to develop a system capable of translating both speech and text between the Tsugaru dialect and standard Japanese.

To address this issue, we are conducting "Hirosaki University × AI × Tsugaru Dialect Project," aiming to develop a bidirectional speech and text translation system between the Tsugaru dialect and standard Japanese using artificial intelligences. This project involves a cross-disciplinary team collecting a wide range of Tsugaru dialect data from various academic and cultural fields for AI training. Additionally, we are systematically organizing the collected dialect data to build a foundational database for the future applications. This paper describes some training methods and their effectiveness of a speech recognition AI using the speech data collected in the project.

Previous attempts at recognizing the Tsugaru dialect using existing tools such as VoiceTra, VOITER, and Google Speech-to-Text have been conducted. However, these tools' dictionaries lack entries for Tsugaru-specific vocabulary, rendering accurate recognition unfeasible. Consequently, there arose a necessity for a speech recognition model trained specifically on Tsugaru dialect words and audio. To address this, we have introduced Wav2Vec2.0[1] which is a speech recognition framework that achieves high accuracy through a combination of pre-training on large-scale unlabeled audio data and fine-tuning with labeled data. Previous research has demonstrated that fine-tuning with a small amount of fixed-phrase data enables the learning of the Tsugaru dialect[2]. In this study, we evaluate the performance of a speech recognition model trained not only on fixed-phrase data but also on free-form conversational data, as well as the impact of using different pretrained models.

II. Evaluation Method

In this research, some models available on Hugging Face(Model A)[3], which are widely used for natural language processing and machine learning, are selected for their suitability for small-scale experiments and are used in the Wav2Vec2.0. Fine-tuning is performed using the base models available on Hugging Face, in addition to the large model(Model B)[4] and multilingual model(Model C)[5] officially provided by Facebook. The training data comprises 5,000 standard Japanese utterances from the JSUT corpus[6] and 530 fixed-phrase utterances and 530 free-form conversational utterances in the Tsugaru dialect. Especially, some of the fixed-phrase data contain sentences with nearly identical content. All reference labels are manually created to align with the actual spoken utterances. All experiments are conducted on a system running AlmaLinux8.9, equipped with an Intel Xeon Gold 5315Y processor, 2 TB of memory, and an NVIDIA RTX6000 Ada GPU. The software environment includes Python 3.11.9, CUDA 12.4, PyTorch 2.5.0.dev20240617, Torchaudio 2.4.0.dev20240617, and Transformers 4.37.2. As preprocessing, the audio data are resampled to a sampling frequency of 16,000 Hz, which is the frequency supported by the above tools. Punctuation marks and symbols are manually removed from the labels to create a character-level vocabulary file.

For test data extraction, the following three types are assumed:

1. 100 randomly selected free-form conversation samples.
2. 100 randomly selected fixed-phrase samples.
3. a combination of 1 and 2.

For training dataset, the following three types are assumed:

a. standard Japanese data combined with free-form conversation data of the Tsugaru dialect excluding the test samples.
b. standard Japanese data combined with fixed-phrase data of the Tsugaru dialect excluding the test samples.
c. a combination of a and b.

Two types of evaluations are conducted, both using a batch size of 8 and a learning rate of 1e-4. The evaluation metric

employed is the Character Error Rate (CER), calculated as the Levenshtein distance divided by the number of characters in the reference label. The Levenshtein distance is defined as the minimum number of single-character edits—insertions, deletions, or substitutions—required to transform the output into the correct label.

(i) Comparison of Training Methods

We retrain Model A using the above training dataset and evaluated them on the corresponding test datasets. The number of training epochs are set from 5 to 45 in increments of 10. For each test dataset, five different random samples are extracted, and training is conducted accordingly.

(ii) Comparison of Pretrained Models

Model A, Model B and Model C are retrain and evaluated using training data c and test data 3. The number of epochs is set from 5 to 45 in increments of 10. Evaluation is conducted across 10 different test data selection patterns.

III. Evaluation Results and Discussion

(i) Comparison of Training Methods

Representative results of CER for 15, 25, and 35 epochs are summarized in TABLE I. Similar trends are observed across other epoch settings. Although no significant reduction in error rates is seen in training and testing on free-form conversation, the error rate stabilized around 80% for both test types. Models trained on fixed phrases achieved low error rates on similar fixed expressions but remained around 90% on free-form conversation inputs. This suggests that training on free-form conversation is more effective for handling such data. In some cases, models trained on free-form conversation produced phonetically similar outputs, indicating potential for improvement by increasing training data and vocabulary coverage.

(ii) Comparison of Pretrained Models

The average results for each model are shown in Fig. 1. The vertical axis represents the error rate, defined as the Levenshtein distance ratio with respect to the correct transcription, while the horizontal axis indicates the number of training epochs. Model A shows an error rate of around 60%, whereas Models B and C achieved lower error rates of approximately 40%, indicating clearly higher accuracy. The output content from Models B and C is also closer to the test data, demonstrating that these models are more effective.

IV. Conclusion

It has been recognized that training on fixed phrases is effective when focusing solely on accuracy for fixed expressions and improving speech recognition accuracy in environments with free-form conversation requires training on non-fixed, natural speech data.

For future improvements in accuracy, either Model B or Model C—both of which have shown high accuracy with minimal difference—will be used for training. We will focus on evaluating how Model C, which has been pre-trained on 53 languages, performs in comparison to Model B. It is also in the scope of our future work to increase the amount of training data and developing models focused on specific tasks with a limited vocabulary scope.

TABLE I
Evaluation results of CER for each dataset and epoch count.

Train/Test	Epoch=15	Epoch=25	Epoch=35
a/1	88.73	84.82	79.85
a/2	80.54	81.43	80.38
a/3	84.01	82.88	80.16
b/1	92.47	90.82	89.77
b/2	52.95	44.26	40.70
b/3	69.75	64.04	61.54
c/1	85.86	80.60	77.59
c/2	54.79	47.05	41.52
c/3	67.99	61.31	56.83

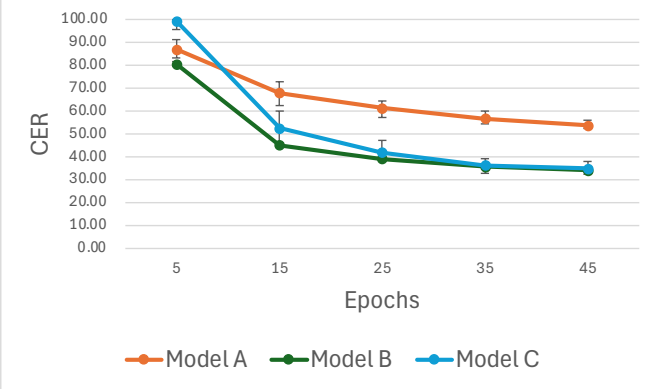

Fig.1. Changes in error rates for each model.

Acknowledgments

This work was partially supported by the Mutsu Ogawara Regional and Industrial Promotion Foundation and JSPS KAKENHI Grant Number JP23H00633 and JP23K25330.

References

[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint*, arXiv:2006.11477, 2020.

[2] Haruto Saito, Akihiro Murakami, and Masashi Imai, "Development of a Tsugaru Dialect Translation System Using a Transparent Display," *DA Symposium 2024 Proceedings*, pp.193–199, Aug., 2024 (in Japanese).

[3] https://huggingface.co/charsiu/zh_w2v2_tiny_fc_10ms

[4] https://huggingface.co/facebook/wav2vec2-large

[5] https://huggingface.co/facebook/wav2vec2-large-xlsr-53

[6] Ryosuke Sonobe, Shinnosuke Takamichi, and Hiroshi Saruwatari, "JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis," *arXiv preprint*, arXiv:1711.00354, 2017.