*Proceedings of the 21st International Conference on Global Research and Education (Inter-Academia 2024)*

# Adapting Large-Scale Pre-trained Models for Unified Dialect Speech Recognition Model

T. Toyama, A. Kai*, Y. Kamiya and N. Takahashi

*Graduate School of Integrated Science and Technology, Shizuoka University 3-5-1 Johoku, Chuo-ku, Hamamatsu, Shizuoka, Japan*

Doi: 10.12693/APhysPolA.146.413

*e-mail: kai.atsuhiko@shizuoka.ac.jp

Recent advancements in deep learning techniques utilizing large-scale data, such as self-supervised learning, have significantly improved the accuracy of speech and language processing technologies for major world languages. However, for dialects with limited transcription resources, technologies like automatic speech recognition and search have yet to be realized at a practical level. This issue is particularly pronounced in Japanese dialects, which are classified into dozens of different and mixed dialects, and remains unresolved. In this study, we focus on two large-scale pre-trained models that have demonstrated top-tier performance in recent automatic speech recognition system research, and present examples of unified automatic speech recognition systems adapted for Japanese dialects, as well as the potential applications of the content detection task — query-by-example spoken term detection. Both compared models are trained on thousands or more hours of multilingual speech, with one being an automatic speech recognition model based on self-supervised learning and the other (Whisper) a model based on multi-task learning, including machine translation. Experiments on automatic speech recognition models are conducted using several tens of hours of adaptation data for both standard Japanese and Japanese dialects, which have distinct characteristics depending on the region. The result shows that the dialect-independent automatic speech recognition model based on the self-supervised learning pre-trained model and 3-step adaptation strategy achieves the best accuracy with a character error rate of 29.2%, suggesting that it is important to consider regional identity due to the diversity and limited resources of Japanese dialects.

topics: automatic speech recognition (ASR), Japanese dialects, large-scale pre-trained models, unified model for dialects

## 1. Introduction

In Japan, the number of dialect speakers is decreasing due to population decline and aging. Therefore, the importance of technology in preserving dialects as cultural and linguistic resources in a form that can be utilized for language analysis and information retrieval, such as in text materials, is increasing. It is possible to use an *automatic speech recognition* (ASR) model to create text materials from dialect speech. However, due to the lack of available dialect corpora for training, the performance of ASR for dialects significantly decreases compared to standard Japanese.

One potential solution to this problem is to construct an ASR system utilizing large-scale pre-trained models for speech processing. *Self-supervised learning* (SSL) is a technique that allows learning latent representations from large amount of unlabeled data and is applied for state-of-the-art speech processing applications [1–3]. It is known that SSL models pre-trained on large unlabeled datasets can improve performance in downstream tasks such as ASR by fine-tuning with only a small amount of labeled data. Our previous studies [4] have shown that by performing multi-task learning of dialect identification and ASR on a large-scale multilingual SSL models, ASR performance can be significantly improved. In contrast to the SSL-based models, Whisper [5] is a multi-task model that has been trained with a large-scale multilingual labeled speech, including low-quality labels, using multi-task learning for tasks such as ASR, machine translation, and language identification, and it is known to exhibit top-tier performance for major languages worldwide. For Japanese, it has been shown to demonstrate high accuracy for standard Japanese (SJ) through fine-tuning, but its adaptability to the diverse Japanese dialects (JD), which vary significantly by region, has not been determined.

This study focuses on an effective method of building a unified ASR model for various Japanese dialects, considering a situation where only a

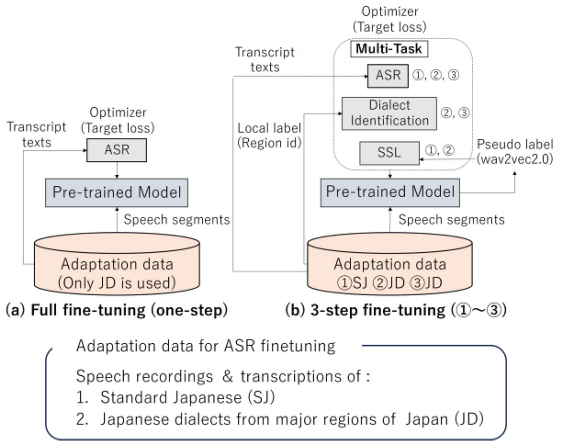
Optimizer
(Target loss)
Multi-Task
Transcript texts
ASR ①, ②, ③
Dialect Identification ②, ③
SSL ①, ②
Local label (Region id)
Pseudo label (wav2vec2.0)
Pre-trained Model
Speech segments
Adaptation data ①SJ ②JD ③JD
Transcript texts
Optimizer (Target loss)
ASR
Pre-trained Model
Speech segments
Adaptation data (Only JD is used)
(a) Full fine-tuning (one-step)
(b) 3-step fine-tuning (①~③)
Adaptation data for ASR finetuning
Speech recordings & transcriptions of :
1. Standard Japanese (SJ)
2. Japanese dialects from major regions of Japan (JD)

Fig. 1. Two adaptive approaches to unified ASR modeling of Japanese dialects.

limited amount of speech and transcripts are available for each region with large differences in dialect characteristics. We compare and analyze the performance of SSL-based ASR models with different adaptation strategies [4] and fine-tuned ASR models based on Whisper, thereby clarifying the current status and challenges of each model in processing dialectal speech. Furthermore, an experimental result on the *query-by-example spoken term detection* (QbE-STD) systems utilizing either the text output from the ASR model or the speech feature from the SSL model is presented to demonstrate the applicability of the proposed models to search tasks for dialects.

## 2. Methods

The approaches used to adapt a large-scale pretraining model for ASR of various dialects are shown in Fig. 1. The following subsections describe the two large-scale pre-trained models that will be compared.

### 2.1. ASR model based on SSL: XLSR and XLS-R

XLSR [6] is a model based on the SSL model wav2vec 2.0 [1] that was pre-trained using 56000 h of unlabeled speech data across 53 languages. To build a model for an ASR task, the most simple approach is full fine-tuning that involves additional training of a pre-trained model using only a small amount of target domain data with target ASR label (Fig. 1a). In a prior study [4], a method was proposed to improve ASR performance for Japanese dialects by applying multi-step adaptation through multi-task learning (Fig. 1b) (hereafter referred to as 3-step fine-tuning). This method minimizes the SSL loss, the *connectionist temporal classification* (CTC) loss for evaluating the alignment with transcripts, and the cross-entropy loss for a *dialect identification* (DID) task that identifies the region where a given dialectal speech was spoken. This method addresses the issue of insufficient training data by leveraging both standard Japanese and dialect data for training. In this study, we use XLS-R [7], which extends the SSL training data to 436000 h across 128 languages, as a comparative model. The SSL models are adapted to build a unified ASR model for Japanese dialects using the finetuning approaches described above, enhanced with our improved adapter-based learning method [8].

### 2.2. Whisper

Whisper is an open-source ASR model pretrained through weakly supervised learning [5]. Weakly supervised learning is a method that trains a model using a large amount of labeled data, including low-quality labels that are mechanically annotated. Whisper has been trained using 680000 h of multilingual audio collected from the web, not only for ASR but also for tasks such as translation, voice activity detection, alignment, and language identification through multi-task learning. Whisper comes in five models with different numbers of parameters: tiny, base, small, medium, and large. In this study, we use the Whisper medium model to build a unified ASR model that is full fine-tuned using only dialect data.

## 3. Experiments: ASR

### 3.1. Datasets

As the standard Japanese speech corpus, we use the *Corpus of Spontaneous Japanese* (CSJ) [9]. This corpus consists of audio recorded in a clean environment, and the transcription format is standardized. In the experiments, we used the katakana transcriptions. For the adaptation training of the standard Japanese, we utilized 61.4 h of monologue lecture and 6 h as validation data. The eval1 test dataset from CSJ was used for evaluation. As for the dialect speech corpus, we used the *Corpus of Japanese Dialects* (COJADS) [10]. This corpus is the largest in Japan, consisting of discourse-formatted speech recorded in real-world environments, including natural dialectal speech from across Japan. The transcriptions are available only in katakana. For the adaptation training of the model, we used 61.5 h of training data and 0.8 h of validation data. For the evaluation data, we used 0.8 h of audio, consisting of 1962 utterances, with no overlap between the training/validation data and the speakers.

### 3.2. Implementation details

Experiments related to XLSR and XSL-R were conducted using the fairseq toolkit [11]. For both models, we used the *large* model, which has 7 feature encoder layers and 24 transformer encoder layers. The hyperparameters were set similarly to the conditions in prior research [4]. Experiments related to Whisper medium were conducted using the Espnet2 [12]. The model consists of 12 transformer encoder layers and 12 transformer decoder layers.

### 3.3. Results

In Table I, we compare the ASR accuracy of each model on standard Japanese (here for CSJ) and Japanese dialects (here for COJADS) using *character error rate* (CER). For standard Japanese ASR, the CER of the SSL models is around 6%, while the Whisper model achieves the highest accuracy with a CER of 4.1%. This result is likely due to the large amount of Japanese included in the pre-training data. In the pre-training of SSL models, XLSR includes 2 h of Japanese, and XLS-R includes 49 h, whereas Whisper uses 7054 h of labeled Japanese speech data for ASR task pre-training. Therefore, compared to the SSL models, the Whisper model is thought to have acquired more acoustic and linguistic knowledge of standard Japanese during the pre-training phase.

For dialectal speech, the Whisper model achieves a CER of 32.9%, which is comparable to the results of the full fine-tuned XLS-R. This suggests that full fine-tuning alone is insufficient for improving ASR performance for dialectal speech when only a small amount of adaptation data is available. In the SSL models, the ASR model using 3-step fine-tuning, which takes into account regional dialectal differences, shows an improvement of about 3–4%.

### 3.4. Analysis

Table II shows a breakdown of the amount of adaptation data and ASR accuracy of XLS-R and Whisper for dialectal speech across eight regions. From the perspective of CER, in full fine-tuning with dialectal speech, Whisper outperforms XLS-R in six out of the eight regions. However, in the case of XLS-R with 3-step fine-tuning, it outperforms Whisper in seven out of the eight regions. When considering the *character error rate reduction* (CERR) from the standard Japanese model CER, XLS-R exhibits a higher reduction rate in five out of the eight regions compared to Whisper in full fine-tuning with dialectal speech. This suggests that XLS-R has a higher adaptability when performing full fine-tuning with a small amount of adaptation data.

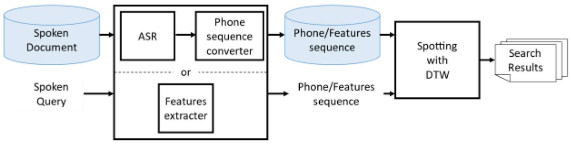

Fig. 2. Overview of the QbE-STD system.

TABLE I

Comparison of ASR accuracy of standard Japanese-adapted model (for CSJ) and unified Japanese dialects-adapted model (for COJADS).

| Pre-trained ASR/SSL model | ASR Adaptation | CER [%] | |
|---|---|---|---|
| | | CSJ | COJADS |
| Target speech: Standard Japanese (SJ) | | | |
| Whisper medium | full fine-tuning | 4.1 | 49.2 |
| XLSR | full fine-tuning | 6.5 | 52.7 |
| XLS-R | full fine-tuning | 6.1 | 51.3 |
| Target speech: Standard dialects (SD) | | | |
| Whisper medium | full fine-tuning | – | 32.9 |
| XLSR | full fine-tuning | – | 34.1 |
| | 3-steps | – | 30.0 |
| XLS-R | full fine-tuning | – | 32.6 |
| | 3-steps | – | 29.2 |

Moreover, the XLS-R model with 3-step fine-tuning outperforms both full fine-tuned models in all regions for dialectal speech. This result indicates that 3-step fine-tuning, which optimizes the model by performing mixed adaptation learning on standard Japanese and dialects and further tailoring to each region, effectively addresses the issue of insufficient training data for dialectal speech, leading to performance improvements across multiple regions regardless of the amount of adaptation data.

## 4. Experiments: QbE-STD

As an application of dialect ASR, one notable example is the *spoken term detection* (STD) task, which involves using a spoken query to identify where the query is spoken within a long recording. The overview of compared STD methods is illustrated in Fig. 2. First method employs the simplest way to implement STD by converting both the target audio and the query into text using an ASR system and then performing spotting through *dynamic time warping* (DTW) between the texts. The second method involves searching directly between the speech features of the target and the query instead of converting them to text. The latter method is often used for low-resource languages.

TABLE II
Breakdown of CERs of unified Japanese dialects-adapted ASR models by geographic region (CERR values represent CER reduction rates to CERs of standard Japanese-adapted ASR models).

| Geographic region in Japan | Adaptation data size [h] | CER (CERR) [%] | | |
|---|---|---|---|---|
| | | Whisper (full) | XLS-R (full) | XLS-R (3-steps) |
| Tohoku | 7.6 | 29.7 (41.8) | 29.9 (43.3) | 24.2 (54.1) |
| Kanto | 17.1 | 25.9 (40.3) | 28.2 (38.3) | 24.7 (46.0) |
| Chubu | 7.9 | 40.9 (34.2) | 44.5 (28.9) | 38.9 (37.9) |
| Hokuriku | 2.4 | 31.4 (34.3) | 29.9 (39.6) | 26.7 (46.1) |
| Kinki | 11.1 | 34.2 (34.2) | 35.2 (36.2) | 32.2 (41.7) |
| Chugoku | 1.8 | 24.0 (39.1) | 26.6 (40.5) | 23.3 (47.9) |
| Shikoku | 1.9 | 15.8 (55.1) | 19.1 (51.0) | 16.0 (59.0) |
| Kyusyu | 11.7 | 43.1 (18.8) | 39.0 (27.1) | 36.1 (32.5) |

TABLE III
Comparison of QbE-STD performance on standard JAPANESE (CSJ) and dialects speech (COJADS).

| Front–end | Method | ASR adaptation | MTWV | |
|---|---|---|---|---|
| | | | CSJ | COJADS |
| Features | filter bank | – | 0.128 | 0.089 |
| | XLS-R | – | 0.443 | 0.269 |
| ASR | Whisper medium | full fine-tuning | 0.766 | 0.553 |
| | XLS-R | 3-steps fine-tuning | – | 0.630 |

### 4.1. Experiment details

In the standard Japanese experiments, we used the audio from the CSJ CORE set (177 lectures) as the search target. For query audio, we automatically generated 923 queries from the CSJ CORE set using a statistics-based method. In the Japanese dialect, we used evaluation data (1962 utterances) from COJADS, which was not used for training the ASR model, as the search target audio and query audio. For query audio, we attempted to extract dialect-specific expressions. From the transcriptions of COJADS, we automatically selected 12 queries as statistically dialect-specific phrases using a SentencePiece [13] tokenizer and document frequency. The STD experiments were conducted using the S3PRL toolkit [14]. The ASR models were the same as those described in Sect. 3. For the feature-based method, the SSL model was the pre-trained XLS-R. The evaluation metric was the maximum term weighted value (MTWV) [15].

### 4.2. Results

The results of evaluating the speech search performance of STD systems using various models are shown in Table III. For the standard Japanese ASR model, we used Whisper, and for the unified dialects ASR model, we used XLS-R (3-step fine-tuning) and Whisper. The results indicate that the accuracy of searches using the filter bank, a well-known speech feature for ASR modeling, was the lowest for both standard Japanese and Japanese dialects, showing that searches using SSL are more effective. Among searches using SSL, the highest accuracy for both standard Japanese and Japanese dialects was achieved with the ASR system. In the Whisper-based standard Japanese ASR system, the CER was 10.4% for documents and 12.6% for queries, whereas in the XLS-R-based unified dialects ASR system, the CER was 29.3% for documents and 15.1% for queries.

Despite the significant difference in CER for documents, similar search accuracy was achieved. The results suggest that using the XLS-R-based unified dialects ASR system has the potential to achieve accuracy close to that of standard Japanese searches, at least for known words. Additionally, the Whisper-based unified dialect ASR system had a CER of 33.9% for documents and 22.6% for queries, showing a decrease in search accuracy compared to XLS-R.

Table IV shows examples of ASR results for dialect phrases. It can be seen that Whisper produces recognition results closer to standard Japanese compared to XLS-R. The difference between adapted ASR models may be attributed to the different pre-training methods and fine-tuning strategies used for diverse target dialects. Whisper uses a larger amount of labeled data of standard Japanese for pre-training, however, no region information is used for fine-tuning. On the other hand, while using

TABLE IV

Examples of dialect phrase recognition.

| Method | Phrase | CER [%] |
|---|---|---|
| | Example 1 | |
| REF | g o z a r i s u | |
| ST* | ございます | |
| Whisper | g o z a m a s u | 37.5 |
| XLS-R | g o z a r i s u | 0.0 |
| | Example 2 | |
| REF | m a d a o r e sh i t a g o d o n a e N d a d e b a y a: | |
| ST* | まだしたことないんだってばさ。 | |
| Whisper | m a d a h o r e s u t a g o d o n e: N d a d e b a y a: | 18.5 |
| XLS-R | m a d a o r e k i t a g o d o n a e N d a d e b a y a: | 3.7 |

*Standard Japanese translation

a smaller amount of data for standard Japanese, the 3-step fine-tuning of the SSL model incorporated region information. This is considered to be one of the reasons why, as shown in Table II, Whisper has a lower CERR in the Kyushu region, which includes dialects with linguistic features that significantly differ from standard Japanese.

## 5. Conclusions

In this study, we built ASR systems based on different pre-trained models and compared their performance. The results showed that Whisper exhibited superior performance for standard Japanese, while multilingual SSL-based ASR fine-tuning demonstrated better performance for dialects. Specifically, the method of simultaneously learning dialect identification and ASR using XLS-R models was shown to be an effective approach to improving the recognition performance of Japanese dialects. Furthermore, the dialect speech search experiments confirmed that for known word queries, the current ASR accuracy can achieve search performance close to that of standard Japanese. Future challenges include the introduction of the idea of 3-step fine-tuning method for the Whisper model in ASR, investigating the search performance with unknown word queries and an increased number of queries for STD, and exploring speech search utilizing intermediate features from ASR models.

## Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP24K00450. The authors would like to thank Shogo Miwa and Raufun Nahar for their help with ASR experiments.

## References

[1] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, in: *Advances in Neural Information Processing Systems*, Vol. 33, Eds. H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin, Curran Associates, 2020 p. 12449.

[2] W.-N. Hsu, B. Bolte, Y.-H.H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed in: *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 29, IEEE, 2021, p. 3451.

[3] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, in: *2022 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2022)*, IEEE, 2022, p. 6152.

[4] S. Miwa, A. Kai. in: *Proc. Interspeech 2023*, 2023, p. 4928.

[5] A. Radford, J.W. Kim, T. Xu, G. Brockman, C. Mcleavey, I. Sutskever, in: *Int. Conf. on Machine Learning*, 2023, p. 28492.

[6] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, M. Auli, in: *Proc. Interspeech 2021*, 2021, p. 2426.

[7] A. Babu, C. Wang, A. Tjandra et al., in: *Proc. Interspeech 2022*, 2022, p. 2278.

[8] N. Takahashi S. Miwa, Y. Kamiya, T. Toyama, R. Nahar, A. Kai, in: *2024 IEEE 13th Global Conf. on Consumer Electronics (GCCE 2024)*, 2024.

[9] K. Maekawa H. Koiso, S. Furui, H. Isahara, in: *Proc. of the 2nd Int. Conf. on Language Resources and Evaluation (LRE'00)*, ELRA, 2000.

[10] COJADS, 2024.

[11] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, M. Auli, in: *Proc. 2019 Conf. of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2019) — Demonstrations*, Association for Computational Linguistics, 2019 p. 48.

[12] S. Watanabe, T. Hori, S. Karita et al., in: *Proc. Interspeech 2021, Hyderabad, India*, 2021, p. 2207.

[13] T. Kudo, J. Richardson, in: *Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, Eds. E. Blanco, W. Lu, 2018, p. 66.

[14] S.-W. Yang, P.-H. Chi, Y.-S. Chuang et al., in: *Proc. Interspeech 2021*, 2021, p. 1194.

[15] Semantic Scholar, "OpenKWS 13 Keyword Search Evaluation Plan 1", 2013.