

Adaptive Method for Self-Supervised Learning Models on Automatic Dialect Speech Recognition Based on Shared Knowledge of Japanese Dialects and Standard Japanese

Naoru Asakawa, Naoki Takahashi, Atsuhiko Kai, Seiichi Nakagawa

Shizuoka University, Hamamatsu, Shizuoka, Japan

{asakawa.naoru.21, takahashi.naoki.20, kai.atsuhiko, nakagawa.seiichi}@shizuoka.ac.jp

Abstract

Speech recognition for Japanese dialects is challenging, and recognition accuracy tends to be lower compared to standard Japanese. Previous research proposed a three-step learning method based on the self-supervised learning (SSL) model XLS-R as the base model, incorporating three multi-task learning tasks: SSL, ASR, and dialect identification (DID). While this achieved improved recognition performance for dialect speech, it faced the issue of degraded recognition performance for standard Japanese. This study proposes an adaptation method to construct a single speech recognition model, based on the prior model, that is suitable for both Japanese dialects and standard Japanese. We explored the use of diverse speech corpora, including ReasonSpeech based on TV broadcast audio and CEJC based on everyday conversational speech, in addition to the standard Japanese speech corpus CSJ and the dialect speech corpus COJADS used in prior research, aiming for knowledge sharing between dialects and standard Japanese. As a result, we confirmed improved recognition performance for both dialects and standard Japanese by including both in the final step of a three-step learning method. We also examined the impact of differences in corpus type and domain on recognition performance.

Keywords: automatic speech recognition, dialect identification, Japanese dialects, self-supervised learning model

1. Introduction

In recent years, research on automatic speech recognition (ASR) systems utilizing large-scale pre-trained models based on self-supervised learning (SSL) has been actively pursued. SSL is a method for acquiring latent audio representations from unlabeled speech data. SSL models such as wav2vec2.0 (Baevski et al., 2020) are known to achieve high performance when fine-tuned on a small amount of labeled data for tasks like ASR. In particular, XLS-R (Babu et al., 2021) is a pre-trained speech foundation model based on wav2vec2.0, trained on large-scale multilingual speech data, demonstrating excellent performance across various language tasks.

Recent studies across languages such as Norwegian, German, and Chinese have emphasized the importance of dialect-aware ASR systems. (Parsons et al., 2025) showed that prosodic features, particularly pitch dynamics, play a crucial role in dialect classification. Their work demonstrated that fine-tuning Whisper models on low-pass filtered audio enhances sensitivity to prosodic contours, improving classification accuracy even on unmodified speech. Similarly, (Blaschke et al., 2025) introduced a benchmark dataset for German dialects and revealed that Whisper models tend to normalize dialectal input toward standard German, highlighting the challenge of preserving dialectal variation. (Xu et al., 2025) proposed a multi-stage training strategy

combining SSL and large language models (LLMs) for Chinese dialect ASR, achieving state-of-the-art performance on multiple dialect datasets.

The number of dialect speakers in Japan is aging and declining, leading to the loss of valuable cultural and linguistic resources. Therefore, research on dialect speech recognition technology has the potential to contribute to the preservation of linguistic resources and the promotion of regional culture.

(Miwa and Kai, 2023; Kamiya et al., 2024) proposed a three-step fine-tuning method for Japanese dialect ASR using XLS-R as the base model. By incorporating three learning tasks—SSL task, ASR task, and dialect identification (DID) task—along with standard Japanese speech, they demonstrated improved recognition performance for dialect speech. However, adapting to dialect speech led to a decline in recognition performance for standard Japanese speech.

Inspired by these cross-linguistic approaches, this paper proposes an adaptation method using the multilingual self-supervised learning model XLS-R to construct accurate speech recognition models for both Japanese dialects and standard Japanese. To build a universal ASR model through knowledge sharing between dialects and standard language, we examined methods for utilizing diverse speech corpus data. This includes not only Japanese dialect speech corpus and standard-language speech corpus used in prior research, but also large-scale speech corpus col-

lected from TV audio and speech corpus primarily consisting of everyday conversations.

(Takahashi et al., 2024) attempted to build a Japanese dialect ASR model using the pre-trained model Whisper (Radford et al., 2022), which achieves state-of-the-art performance for many major languages, to build a Japanese dialect ASR model, and compared it with XLS-R-based models. However, its recognition performance for dialect speech was inferior to the XLS-R-based models by (Miwa and Kai, 2023; Kamiya et al., 2024), leaving challenges in refining adaptation methods for dialect speech.

This paper describes an adaptation method using the multilingual self-supervised learning model XLS-R to construct accurate speech recognition models for both Japanese dialects and standard Japanese. To build a universal ASR model through knowledge sharing between dialects and standard language, we examined methods for utilizing diverse speech corpus data. This includes not only Japanese dialect speech corpus and standard-language speech corpus used in prior research, but also large-scale speech corpus collected from TV audio and speech corpus primarily consisting of everyday conversations.

2. Related Work

2.1. Self-Supervised and Multilingual Models for Dialect Speech Recognition

XLS-R (Babu et al., 2021) is a model that learns audio representations based on the SSL architecture wav2vec2.0 (Baevski et al., 2020), trained on multilingual audio data comprising 128 languages and 436,000 hours. By fine-tuning with a small amount of labeled data, it can be applied to various speech tasks such as speech recognition, translation, and speaker identification. Furthermore, it is known to be particularly effective for tasks involving low-resource languages due to the rich knowledge it acquires across languages.

Recent work by (Parsons et al., 2025) investigated the role of prosody in dialect identification using Whisper embeddings. Their experiments showed that models trained on low-pass filtered audio, which emphasizes pitch contours, outperform those trained on unmodified or monotonized audio. This suggests that prosodic features are computationally exploitable for dialect modeling and may be beneficial for Japanese dialect ASR as well.

(Blaschke et al., 2025) introduced a multi-dialectal dataset for German ASR and dialect-to-standard speech translation. Their analysis revealed that Whisper models often produce outputs

closer to standard German, even when trained on dialectal input. This highlights the challenge of preserving dialectal variation in ASR outputs and the need for models that can flexibly handle both dialectal and standardized forms.

(Xu et al., 2025) proposed a four-stage training strategy combining SSL-based speech encoders and LLMs for Chinese dialect ASR. By pretraining on 300,000 hours of unlabeled dialectal speech and fine-tuning with 40,000 hours of labeled data, they achieved state-of-the-art performance. Their findings underscore the effectiveness of combining SSL models with LLMs and multi-stage training for low-resource dialect scenarios, which may inform future Japanese dialect ASR research.

2.2. Multi-step Fine-Tuning for Japanese Dialects

(Miwa and Kai, 2023; Kamiya et al., 2024) applied the aforementioned SSL model to ASR tasks for various dialects. Using both standard Japanese speech (CSJ: Corpus of Spontaneous Japanese) and dialect speech (COJADS: Corpus Of Japanese Dialect Speech), they proposed a three-step fine-tuning method employing self-supervised learning (SSL), automatic speech recognition (ASR), and dialect identification (DID) across three learning tasks and two adapters. (Kamiya et al., 2024)

Their approach demonstrated improved recognition performance for dialect speech, but also revealed a decline in recognition performance for standard Japanese, indicating a loss of standard-language knowledge.

Building on these insights and the findings from cross-linguistic studies, our work explores how diverse corpus types and shared training strategies can mitigate performance trade-offs and enable robust recognition across both dialectal and standard Japanese speech.

3. Three-step Fine-Tuning Approach

Figure 1 shows the data used at each step, the learning tasks, and the parameters adjustable during training (red). AdaptS and AdaptD represent adapters for standard Japanese and dialects, respectively.

3.1. 1st Step: SSL + ASR Learning on Standard Japanese Speech

Based on the SSL model XLS-R, the first training stage involves multitask learning of SSL and ASR using standard Japanese speech. During training, only the standard Japanese adapter is trained, while most other model parameters are fixed. This

stage aims to acquire knowledge about standard Japanese speech and adapt the multilingual SSL model for the Japanese ASR task.

3.2. 2nd Step: SSL + ASR + DID Learning on Dialect Speech

In the second stage of training, we perform multi-task learning for SSL, ASR, and DID using dialect speech. During training, only the dialect adapter is trained, while the majority of parameters in the other models are fixed. This stage aims to acquire knowledge about Japanese dialect speech.

3.3. 3rd Step: ASR + DID Learning on Dialect or Dialect and Standard Japanese Speech

In the final training stage, we perform two-task learning for ASR and DID using dialect speech. Training fixes the adapter and trains only the parameters of the Transformer layer on dialect speech.

Table 1 shows the speech recognition performance of the model created by this method for dialect speech and standard Japanese speech. The three-step learning demonstrated a positive effect on dialect recognition performance, but recognition performance for standard language decreased, indicating a loss of standard-language knowledge accompanying adaptation to dialect speech. It should be noted that the difficulty in recognizing dialect speech (COJADS; mainly elderly speakers and free conversation) stems not only from its dialect nature but also from factors such as speakers being elderly and the data being free conversational speech.

3.4. Improved 3rd step

This paper describes efforts to construct a single ASR model adapted to both Japanese dialects and standard Japanese, based on and improved from the three-step learning approach of prior research. Key challenges in the prior research methods include a decline in recognition performance for standard Japanese when adapting to dialects, and significant differences in domains such as utterance styles and linguistic features between COJADS and CSJ. To address these issues, experiments were conducted on three approaches: a method using both standard Japanese and dialect data for training in the final step, a method using ReasonSpeech as the standard Japanese dataset, and a method using CEJC as the standard Japanese dataset (Figure 2).

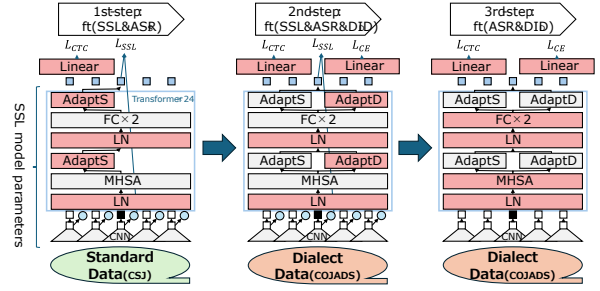


Figure 1: Three-step learning from SSL models for Japanese dialect ASR (Kamiya et al., 2024) (ft (task) denotes fine-tuning on the single or multiple tasks listed in parentheses.)

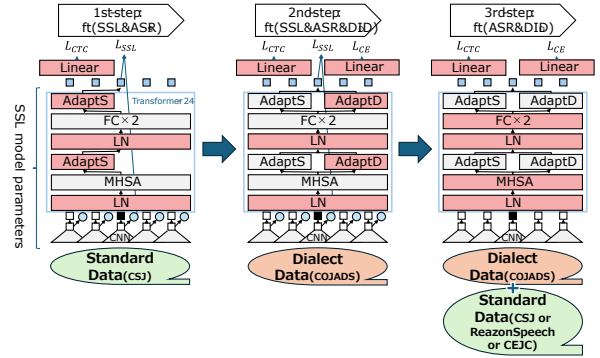


Figure 2: Three-step learning from SSL models for Japanese dialect and Standard Japanese ASR

4. Datasets and Experimental Methods

4.1. Datasets

The dataset used in the experiment is shown in Table 2.

4.1.1. COJADS

Data from both standard Japanese and dialects are used for 3 steps of learning. The Corpus of Japanese Dialects (COJADS) (National Institute for Japanese Language and Linguistics (NINJAL), 2021) was used as the dialect corpus. COJADS is a major Japanese dialect corpus consisting primarily of multi-speaker conversational audio, containing a large amount of natural dialects commonly found in spoken language. It includes examples with low recording quality and transcriptions with low accuracy. Transcriptions are written solely in katakana (grapheme corresponding to syllable). Approximately 62 hours of training data and 0.83 hours of evaluation data were used. In this experiment, the COJADS dataset was divided into 17 regions, and the regional labels were used as supervisory signals for multi-task learning, including

Fine-tuning Method / Datasets	Adapter usage	Training Module			CER[%]	
		1st step:	2nd step:	3rd step:	CSJ	COJADS
One-step / COJADS only	none	-	-	-	15.0	32.6
Three-step / CSJ (1st)+COJADS (2nd&3rd)	none	CNN+Q+TF	CNN+Q+TF	TF	11.1	28.8
Three-step / CSJ (1st)+COJADS (2nd&3rd)	yes	AdaptS	-	-	4.4	48.2
	yes	AdaptS	AdaptD	-	14.2	35.1
	yes	AdaptS	AdaptD	TF	11.2	29.2

Table 1: Comparison of speech recognition accuracy for XLS-R models trained with multi-step learning for Japanese dialect ASR tasks (Kamiya et al., 2024). (CER: Character Error Rate, Q: Quantizer, TF: Transformer module, Adapter internal output dimension 1024)

	Training			Validation			Testing		
Dataset	Utterances	Speakers	Time[h]	Utterances	Speakers	Time[h]	Utterances	Speakers	Time[h]
CSJ	151783	918	227.4	3590	30	5.82	1272	10	1.83
COJADS	141425	304	61.5	2033	231	0.86	1962	36	0.83
Reazon	152102	-	261.6	8463	-	14.5	-	-	-
CEJC	418973	918	141.5	52755	876	17.7	-	-	-

Table 2: Dataset statistics for dialect and standard Japanese ASR training

	Training			Validation			Testing		
Region	Utterances	Speakers	Time[m]	Utterances	Speakers	Time[m]	Utterances	Speakers	Time[m]
0	1057	2	28.0	8	2	0.2	75	1	1.7
1	7305	22	191.9	154	14	3.8	171	3	4.4
2	9083	20	237.7	85	16	2.1	91	2	3.2
3	4115	10	99.9	60	9	1.4	62	1	1.2
4	35335	59	926.6	570	47	14.9	247	3	6.1
5	1145	4	26.0	10	2	0.2	59	1	1.4
6	6844	10	142.5	93	8	2.0	68	1	1.1
7	10854	10	306.1	163	8	4.4	59	3	1.3
8	5566	20	143.2	68	15	2.0	112	3	2.7
9	25874	61	664.9	299	38	7.4	247	7	5.6
10	3012	7	81.0	44	6	1.0	175	2	4.3
11	999	5	27.4	15	3	0.3	52	1	2.5
12	4296	12	113.8	68	10	1.8	98	1	2.3
13	14465	29	378.6	222	23	5.4	90	2	3.1
14	9141	22	259.2	132	20	3.5	66	2	1.5
15	651	2	13.7	12	2	0.3	118	1	3.2
16	1683	9	50.3	30	8	0.8	172	2	4.3

Table 3: Detailed breakdown of each dataset by region for COJADS (Region numbers correspond to those in the map in Appendix A.)

dialect identification (DID). The distribution of the number of utterances, speakers, and total audio duration for each region is shown in Table 3.

4.1.2. CSJ

For the standard language corpus, we used the Corpus of Spontaneous Japanese (CSJ) (Maekawa, 2003). This dataset consists primarily of lecture recordings made in a clean environment, with transcription notation strictly standardized. For this study, we use the katakana transcription to enable simultaneous training and comparison with COJADS. Approximately 230 hours of

academic conference lectures were used as training data, and approximately 6 hours as evaluation data. The evaluation utilized the eval1 dataset defined within the corpus.

4.1.3. ReazonSpeech

In a subset of our experiments, the medium-v1 version of the ReazonSpeech audio corpus (Yin et al., 2023) was used as the standard-language corpus. The ReazonSpeech audio corpus is an audio corpus created from TV recording data and its subtitles, containing diverse speech styles such as news, variety shows, and documentaries. It

can be considered to contain substantial information on dialectal speech features. For this study, the medium-v1 dataset, containing 300 hours of data, was used as the standard-language dataset. Of this, 262 hours were used as training data and 15 hours as validation data. The ReazonSpeech corpus uses subtitles from TV broadcasts as labels, which include hiragana, kanji, symbols, and other characters. For simultaneous training with COJADS, this study required labels using only katakana. Therefore, an ASR model was created through three-step learning using data concatenating all CSJ and COJADS. This model was used to transcribe ReazonSpeech data into kana characters, and the resulting automatic transcription was directly used as labels. Some data contained poorly transcribed examples and such data was excluded from the training set.

4.1.4. CEJC

The Corpus of Everyday Japanese Conversation (CEJC) (Koiso et al., 2022) is an audio corpus primarily focused on everyday conversation. It is an audio corpus targeting conversations that naturally arise from the participants' own motivations and purposes within everyday situations, and it contains a well-balanced collection of conversations from diverse situations involving speakers of various genders and ages. It is also considered to contain a significant amount of information on dialectal speech features. Unlike ReazonSpeech, it allows for the creation of accurate katakana text labels. In this study, approximately 140 hours of CEJC were used as training data, and approximately 18 hours were used as validation data.

4.2. Experimental Methods

4.2.1. Three-step Learning with Standard-Language Speech Data

Previous studies employed a three-step learning process aimed at adapting ASR to Japanese dialects during the final learning step, resulting in degraded recognition performance for standard Japanese (Table 1). To address this issue, we conducted learning using both standard Japanese and dialect data in the final step. Three data combination methods were considered: (1) Using the entire standard language (230h) and dialect (62h) datasets; (2) Allocating equal amount of training time to standard language and dialect (standard language 60h, dialect 62h); (3) Adjusting standard-language data to match the data volume per dialect region (standard language 3.7h, dialect 62h=3.7×17). The combined data was used in the final step of the three-step training process.

4.2.2. Three-step Learning with ReazonSpeech as Standard-Language Speech Data

In the 1st step and the 3rd step, the Corpus of Spontaneous Japanese (CSJ) was used as the standard-language dataset for experiments conducted thus far. This section describes a model created using the newly developed ReazonSpeech audio corpus as the standard-language data, rather than CSJ. ReazonSpeech is an audio corpus utilizing TV broadcast audio and subtitles as data, containing diverse dialectal features from various speakers and data in various speech styles, including dialogic speech. Therefore, for simultaneous learning with COJADS, which primarily uses discourse audio, we considered that using ReazonSpeech as standard-language data would enable more effective learning than using CSJ, which mainly uses lecture audio and only provides dialectal features from a limited set of speakers.

4.2.3. Three-step Learning with CEJC as Standard-Language Speech Data

In the 1st step and the 3rd step, the experiments thus far have used the Corpus of Spontaneous Japanese (CSJ) and ReazonSpeech as standard-language datasets. This section describes a newly created model using the CEJC speech corpus as standard-language data. CEJC is a speech corpus primarily focused on everyday conversation, containing a well-balanced collection of conversations across diverse situations involving multiple speakers of various genders and ages. Speech in multi-person conversational styles is expected to contain diverse dialectal features. Therefore, for simultaneous learning with COJADS, which primarily uses discourse audio, we considered that using CEJC as standard-language data would enable more effective learning than using CSJ, which mainly uses lecture audio and only provides dialectal features from a limited set of speakers. Furthermore, by utilizing the pronunciation data within the CEJC dataset, accurate katakana text labels can be created. This was expected to improve recognition performance compared to using ReazonSpeech.

4.3. Experimental Conditions

Experiments were conducted using the fairseq toolkit (Ott et al., 2019). We employed the wav2vec2.0 model defined as large size, consisting of 7 CNN layers and 24 Transformer layers, and utilized XLS-R pre-trained on 128 languages and 436,000 hours of audio. In experiments using adapters, adapters were inserted into all Transformer layers. The learning rate was set to 3e-

5 when training the Transformer layers and $5e-4$ when not training them. When not including SSL training, the quantizer for extracting quantized representations and the linear layer for the SSL task were removed. The number of updates was performed between 25k and 40k.

5. Results and Discussion

5.1. Three-step Learning with Standard-Language Speech Data

Table 4 shows the results of fine-tuning all parameters and fine-tuning using an adapter. The model that achieved the lowest CER for dialect speech was the one trained using an adapter with 3.7h of standard-language data (CSJ). Using both standard language and dialect simultaneously in the final training step enabled generalization performance for both, and depending on the proportion of training data, recognition performance for each further improved.

Compared to the recognition performance of Japanese dialect ASR models using the Whisper model described in (Takahashi et al., 2024), the model obtained by this method demonstrated superior recognition performance for dialect speech. This is likely because, while the Whisper model tends to misrecognize speech features specific to spoken language—such as fillers and hesitations—and unknown words frequently appearing in dialect speech as existing words, the XLS-R-based model trained to capture these speech features more faithfully.

Table 5 shows the recognition performance for each COJADS dialect region using the proposed model trained with 3.7 hours of standard Japanese data, which achieved the best performance for dialect speech in this study. The table indicates that recognition performance varies across dialect regions, with some regions being recognized with relatively low CER, while others remain difficult to recognize with higher CER. In particular, Region 7 and Region 16 exhibit high CER, suggesting that these dialects may be difficult to recognize due to differences in phonology and vocabulary, as well as the influence of recording conditions. Compared with the baseline model trained only on the CSJ dataset using only one-step ASR training procedure, the proposed method improves CER for all dialect regions. These results indicate that the proposed method consistently improves recognition performance for dialect speech across different regions.

Base model / Fine-tuning (Final step data)	CER [%]	
	CSJ	COJADS
XLS-R/Three-step learning adapter-based (full) fine-tuning (CSJ:0h, COJADS:62h)	11.2 (11.1)	29.2 (28.8)
(CSJ:3.7h, COJADS:62h)	5.5 (6.6)	28.5 (28.8)
(CSJ:60h, COJADS:62h)	4.3 (4.7)	29.2 (28.7)
(CSJ:230h, COJADS:62h)	3.7 (3.7)	29.4 (29.3)
Whisper/One-step learning full fine-tuning (CSJ:0h, COJADS:62h)	10.3	36.1
(CSJ:230h, COJADS:62h)	4.5	34.2

Table 4: Comparison of recognition accuracy when including standard-language learning data in the final step of three-step learning.

Region	Baseline	Proposed model
0	47.4	16.5
1	58.2	29.0
2	45.7	17.6
3	69.7	30.9
4	43.0	23.9
5	56.3	29.1
6	56.6	29.9
7	78.2	59.2
8	50.5	25.0
9	55.3	32.5
10	45.7	23.4
11	43.8	21.0
12	35.9	15.9
13	44.6	24.7
14	56.3	34.3
15	48.8	37.3
16	55.4	40.6
Average	50.0	28.5

Table 5: Breakdown of regional COJADS performance for baseline and proposed XLS-R-based models (CER [%])

5.2. Three-step Learning with ReasonSpeech as Standard-Language Speech Data

The results of experiments using ReasonSpeech are shown in Table 6. Using ReasonSpeech as the standard-language data improved recognition performance for dialect speech after completing the 1st step of three-step fine-tuning compared to conventional methods (in Table 1), indicating high compatibility between ReasonSpeech and COJADS as datasets. However, after completing the third step of learning, recognition performance declined compared to the conventional method using CSJ, even after implementing measures to balance the data time allocation for standard language. One reason for this outcome is the

accuracy issue of the model transcribing ReazonSpeech data into kana characters (ReazonSpeech corpus has not perfect katakana transcription, but only imperfect kanakanji transcription). For this transcription to create kana labels for ReazonSpeech, we used the ASR model currently offering the best CSJ recognition performance within our research methodology. However, since no adaptive learning was performed specifically for ReazonSpeech, the training data contained a significant amount of incorrectly transcribed data. Comparing the kana labels from the ASR model's recognition results with the kana-kanji text labels provided for ReazonSpeech, and evaluating a portion of the dataset, yielded a recognition accuracy of 25.1% CER (Substitution:6.4%, Deletion:5.4%, Insertion:13.3%). Using such inaccurate labels is considered the primary reason recognition accuracy did not improve. If more accurate transcription of kana-kanji labels using kana characters becomes possible, using ReazonSpeech as the standard language in this method is expected to improve recognition performances compared to conventional approaches.

Another issue is that the TV broadcast subtitle data underlying the ReazonSpeech text data sometimes fails to faithfully reflect parts of the spoken content, including speech-specific fillers and hesitations. Meanwhile, the main task here is an ASR task for dialects and standard Japanese using an XLS-R-based model, which requires accurately capturing speech-specific acoustic features. While using the recognition results from the ASR model as labels may have mitigated this mismatch, it has made it difficult to assess the accuracy of the newly assigned kana labels due to these issues. How to incorporate the vast ReazonSpeech dataset for kana learning or dialect learning remains a future challenge.

ASR Learning Method (Fine-tuning Data)	CER [%]	
	CSJ	COJADS
Multi-step adapter-based fine-tuning		
1st step (Reazon:260h)	7.2	40.3
2nd step (COJADS:62h)	15.6	34.6
3rd step		
(Reazon:3.7h, COJADS:62h)	9.0	28.8
(Reazon:260h, COJADS:62h)	6.1	29.4

Table 6: Comparison of recognition accuracy when using ReazonSpeech as standard-language learning data in a three-step learning approach.

5.3. Three-step Learning with CEJC as Standard-Language Speech Data

Table 7 shows the results of experiments using CEJC as the standard-language dataset.

The results indicate that when using CEJC as the standard-language dataset, similar to ReazonSpeech, recognition performance for dialect speech improved after completing the first step of three-step fine-tuning. However, after completing all 3 steps of training, recognition performance decreased compared to the conventional method using CSJ. CEJC was expected to improve recognition performance because, as conversational speech, it has a higher domain match rate with COJADS than ReazonSpeech. Unlike ReazonSpeech, it allows for the creation of accurate katakana text labels, and transcriptions faithfully include fillers and hesitations. However, contrary to expectations, incorporating CEJC into the training data did not yield any noticeable effect. Reasons for this outcome include the relatively small amount of CEJC data, the use of CEJC data as training without accounting for overlapping utterances within the data, and the inclusion of utterances from speakers without dedicated microphones in the training data. Since how data is selected during this preprocessing step significantly impacts recognition performance, it is expected that using more refined CEJC data for training could lead to improved recognition performance compared to this experiment.

ASR Learning Method (Fine-tuning Data)	CER [%]	
	CSJ	COJADS
Multi-step adapter-based fine-tuning		
1st step (CEJC:140h)	9.6	45.3
2nd step (COJADS:62h)	15.6	34.3
3rd step		
(CEJC:3.7h, COJADS:62h)	10.8	29.0
(CEJC:140h, COJADS:62h)	8.8	28.8

Table 7: Comparison of recognition accuracy when using CEJC as standard-language learning data in a three-step learning approach.

6. Conclusion

This paper introduces a method for constructing a speech recognition model adapted to both Japanese dialects and standard Japanese, based on the multilingual self-supervised learning model XLS-R. Compared to previous models that employed multi-task learning including dialect identification and introduced adapters for dialects and standard Japanese to improve learning efficiency, we investigated a method that simultaneously utilizes both dialects and standard Japanese in the final training step. We also explored methods utilizing new standard Japanese datasets: the ReazonSpeech audio corpus and the Corpus of Everyday Japanese Conversation (CEJC). Using both dialect and standard Japanese in training enhanced

the generalization performance of the three-step fine-tuning model. Depending on the proportion of standard Japanese in the training data, we achieved improved recognition performance for both dialect and standard Japanese. Future challenges include exploring methods for accurately transcribing ReasonSpeech data into kana characters, improving preprocessing methods for the CEJC dataset, and investigating ways to incorporate richer datasets into training. Our goal is to further enhance recognition performance for both standard Japanese and dialects.

7. Acknowledgements

This research was supported by a Grants-in-Aid for Scientific Research (Grant Number JP24K00450) from the Japan Society for the Promotion of Science (JSPS).

8. Bibliographical References

Arun Babu, Changan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. [Xls-r: Self-supervised cross-lingual speech representation learning at scale](#).

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. [wav2vec 2.0: A framework for self-supervised learning of speech representations](#). *CoRR*, abs/2006.11477.

Verena Blaschke, Miriam Winkler, Constantin Förster, Gabriele Wenger-Glemser, and Barbara Plank. 2025. [A Multi-Dialectal Dataset for German Dialect ASR and Dialect-to-Standard Speech Translation](#). In *Interspeech 2025*, pages 913–917.

Yuta Kamiya, Shogo Miwa, and Atsuhiko Kai. 2024. [A parameter-efficient multi-step fine-tuning of multilingual and multi-task learning model for japanese dialect speech recognition](#). In *2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6.

Shogo Miwa and Atsuhiko Kai. 2023. [Dialect Speech Recognition Modeling using Corpus of Japanese Dialects and Self-Supervised Learning-based Model XLSR](#). In *Proc. INTERSPEECH 2023*, pages 4928–4932.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. [fairseq: A fast, extensible toolkit for sequence modeling](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Phoebe Parsons, Heming Strømholte Bremnes, Knut Kvale, Torbjørn Svendsen, and Giampiero Salvi. 2025. [Effects of Prosodic Information on Dialect Classification Using Whisper Features](#). In *Interspeech 2025*, pages 2785–2789.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#).

Naoki Takahashi, Shogo Miwa, Yuta Kamiya, Takumi Toyama, Raufun Nahar, and Atsuhiko Kai. 2024. [Comparison of large pre-trained models and adaptation methods for japanese dialects asr](#). In *2024 IEEE 13th Global Conference on Consumer Electronics (GCCE)*, pages 811–814.

Tohoku NLP Group. 2023. [Bert large japanese \(character-level tokenization with whole word masking, cc-100 and jawiki-20230102\)](#). <https://huggingface.co/tohoku-nlp/bert-large-japanese-char-v2>.

Tianyi Xu, Hongjie Chen, Qing Wang, Lv Hang, Jian Kang, Jie Li, Zhennan Lin, Yongxiang Li, and Lei Xie. 2025. [Leveraging LLM and Self-Supervised Training Models for Speech Recognition in Chinese Dialects: A Comparative Analysis](#). In *Interspeech 2025*, pages 584–588.

9. Language Resource References

Hanae Koiso and Haruka Amatani and Yasuharu Den and Yuriko Iseki and Yuichi Ishimoto and Wakako Kashino and Yoshiko Kawabata and Ken'ya Nishikawa and Yayoi Tanaka and Yasuyuki Usuda and Yuka Watanabe. 2022. [Design and Evaluation of the Corpus of Everyday Japanese Conversation](#). European Language Resources Association. PID <https://aclanthology.org/2022.lrec-1.599/>.

Kikuo Maekawa. 2003. [Corpus of spontaneous Japanese: its design and evaluation](#). ISLRN 280-594-494-328-0.

National Institute for Japanese Language and Linguistics (NINJAL). 2021. *Corpus of Japanese Dialects*. PID <https://www2.ninjal.ac.jp/cojads/index.html>.

Yue Yin and Daijiro Mori and Seiji Fujimoto. 2023. *ReazonSpeech: A free and massive corpus for Japanese ASR*. PID https://research.reazon.jp/_static/reazonspeech_nlp2023.pdf.

A. Regional Division of the COJADS Dialect Dataset

In this study, the COJADS dialect data were divided into 17 regions and labeled accordingly. The details of this regional division are shown in Figure 3. The figure visualizes the locations where speech data were recorded for each prefecture in Japan and indicates which dialect region each location corresponds to. These regional labels were used as supervisory labels for DID (Dialect Identification) in the multi-task learning setup.

B. Perplexity of Language Model on Dialect Regions

In this appendix, we show the perplexity of a BERT language model, fine-tuned from the pre-trained bert-large-japanese-char-v2 model (Tohoku NLP Group, 2023) using standard Japanese texts, for texts from each dialect region in COJADS. This allows us to evaluate how well a model trained on standard Japanese adapts to dialectal variations and provides a quantitative view of linguistic differences between standard Japanese and regional dialects.

Table 8 shows that, overall, the perplexity for dialect regions is high when using a language model trained on standard Japanese, indicating that the model is not well adapted to regional dialects. Re-

gions 1 (western Tohoku), 15 (Kagoshima), and 16 (Okinawa) exhibit particularly high perplexity, reflecting substantial differences in vocabulary and expression from standard Japanese. Although the dialect for the region 4 is similar to the standard Japanese, the perplexity is large. It is the reason that the contents of dialect and standard Japanese are elderly people’s daily conversation and academic presentation, respectively. In contrast, the perplexity for the standard Japanese corpus (CSJ) is low, showing that the model performs well on standard Japanese. These results provide a quantitative illustration of the linguistic differences between Japanese dialects and standard Japanese.

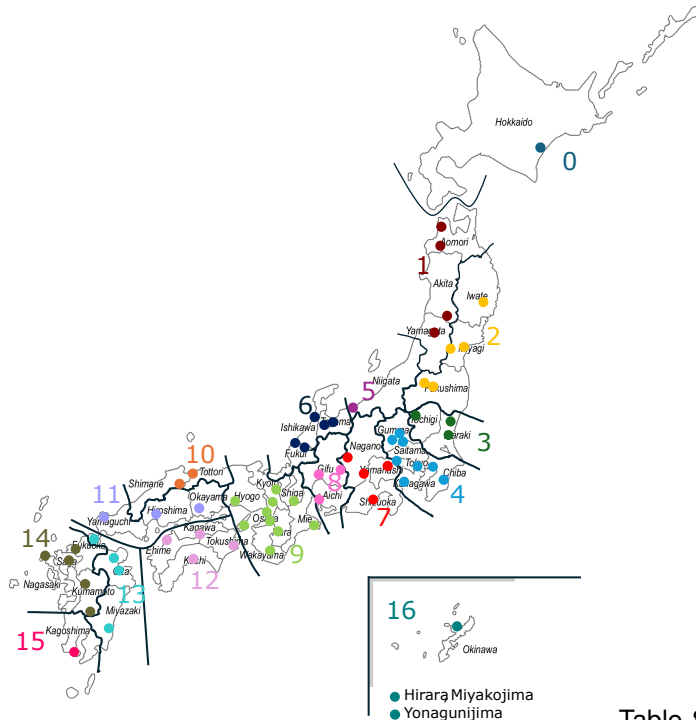


Figure 3: The 17 dialect regions of COJADS and the locations where speech data were recorded

Region	Perplexity
0	88.5
1	146.1
2	91.9
3	60.0
4	60.9
5	75.0
6	76.6
7	92.2
8	68.2
9	84.6
10	58.1
11	62.4
12	99.9
13	84.5
14	75.5
15	140.2
16	143.6
Average	91.4
CSJ	9.5

Table 8: Perplexity of the BERT language model fine-tuned on standard Japanese for each COJADS dialect region and for standard Japanese (CSJ).