INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Dialect Speech Recognition Modeling using Corpus of Japanese Dialects and Self-Supervised Learning-based Model XLSR

Shogo Miwa, Atsuhiko Kai

Graduate School of Integrated Science and Technology, Shizuoka University, Japan
{miwa.shogo.18,kai.atsuhiko}@shizuoka.ac.jp

Abstract

In order to utilize the large amount of historical speech resources for applications such as linguistic analysis and retrieval, automatic speech recognition technology that can handle a variety of dialects is required. Although there are many dialects in the Japanese language, there have been no reports of speech recognition models that cover almost all Japanese dialects using only shared dialect resources. This paper presents a baseline for dialect speech recognition of spoken Japanese using a nationwide corpus of Japanese dialects released in 2022. Specifically, the paper presents results on: 1) the effectiveness of adapting a self-supervised learning model, which has been shown to be effective for low-resource languages, to the dialect corpus; 2) the effectiveness of combining both automatic speech recognition and dialect region identification tasks, or when used in conjunction with a large-scale corpus of standard Japanese, within the framework of self-supervised learning.

Index Terms: automatic speech recognition, dialect identification, wav2vec2.0, Corpus of Japanese Dialects

1. Introduction

The number of Japanese dialect speakers is declining and a valuable cultural and linguistic resources are being lost. While the construction of textual materials using high-performance automatic speech recognition (ASR) models is effective for standard Japanese, the performance of speech recognition for dialects that are linguistically and acoustically different from standard Japanese is greatly degraded due to the lack of dialect corpus for training. Although there have been studies using dialect speech recognition models based on handmade dialect corpus [1, 2], creating a handmade dialect corpus is very costly. However, with the release of the Corpus of Japanese Dialects (COJADS) [3] in 2022, which includes speech data from all prefectures of Japan and spans over 60 hours, it is now possible to use publicly available large-scale dialect corpus material for research purposes. COJADS is an extremely realistic dialect corpus that includes natural dialects that are more likely to appear in spoken Japanese in natural condition, since many of the recordings are of low quality and are spoken by several elderly people in a conversational style. Because it was not developed for speech recognition, the transcriptions are in katakana only, and do not include information on kanji or phonemes. This paper reports on an attempt to build an universal speech recognition model for various dialects throughout Japan using this dialect corpus. Self-supervised learning (SSL) [4, 5, 6, 7, 8], which has been shown to be effective for low-resource languages, is used in this research. Self-supervised learning is a method for learning potential speech representations from speech alone, and it is known to improve performance when the learned speech representations are used for various downstream tasks such as ASR [9, 10, 11, 12, 13, 14]. XLSR [15] is one of the SSL model built on wav2vec2.0 [6], with speech data of 53 different languages, and has shown to be effective in low-resource language applications. In this study, the XLSR model is used as the basis for fine-tuning to improve the performance of dialect speech recognition. We show that the framework of self-supervised learning adapted to dialect corpus is effective, when combined with both automatic speech recognition and dialect identification (DID) tasks, or when used in conjunction with a large corpus of standard Japanese. The proposed method obtained relative character error rate reductions of up to 8.9% from models when the ASRs models ware simply fine-tuned.

2. Related work

In a previous study of ASR for Japanese dialects [2], the authors improved ASR performance by multitask learning with DID using a handmade dialect corpus. For low-resource languages, ASR models using XLSR, a self-supervised learning model trained on multiple languages, have shown state-of-the-art performance [15, 16]. In [17, 18, 19], the authors adapt phonetic representation and transcription knowledge from the source language to the target language by performing multitask learning of SSL and ASR. With these as a reference, we compare the multitask learning methods of SSL, ASR, and DID using XLSR, and examine the effectiveness of adaptation from standard Japanese to a universal ASR model for various dialects.

3. Methods

In this study, experiments were conducted using wav2vec2.0 [6], a self-supervised learning model, and the Hybrid-CTC/Transformer ASR model [20, 21, 22] as a comparison.

3.1. Hybrid-CTC/Transformer ASR

Hybrid-CTC/Transformer is a combination of CTC (Connectionist Temporal Classification) [23] and Transformer encoder-decoder model. Raw audio $\mathbf{X}$ is the input to the encoder layer, which outputs an intermediate representation $\mathbf{H}$. The $\mathbf{H}$ is then fed as input to the linear layer, which calculates the CTC loss L_{CTC}. The decoder layer, on the other hand, takes $\mathbf{H}$ and the past predicted sequence $\mathbf{Y}'$ as input and calculates the Decoder loss $L_{Decoder}$. The Decoder outputs one character at a time in an autoregressive fashion. The CTC loss and Decoder loss are weighted and added together to obtain the overall loss L.

$$L = \lambda L_{CTC} + (1 - \lambda) L_{Decoder} \tag{1}$$

where λ is an adjustable parameter.

10.21437/Interspeech.2023-2463

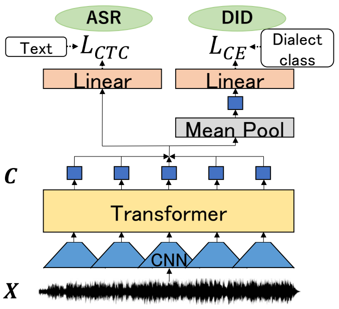

Figure 1: *ASR&DID multitask learning model*

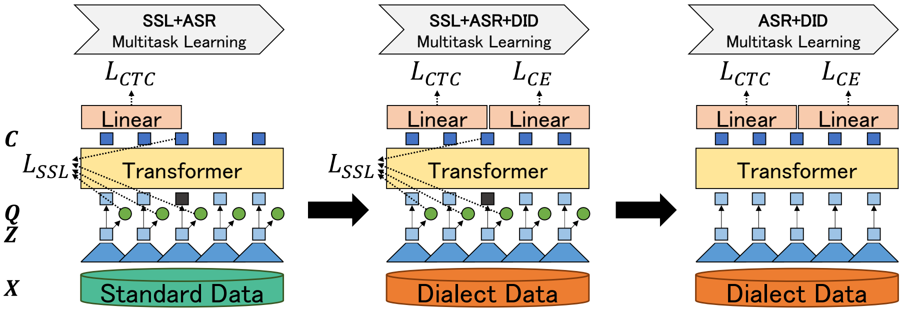

Figure 2: *Adaptation flow from standard Japanese speech to Japanese dialects. The final model is called "XLSR+ft(SSL&ASR)$_S$+ft(SSL&ASR&DID)$_D$+ft(ASR&DID)$_D$".(ft: fine-tuning, CSJ(S):standard Japanese speech Corpus, COJADS(D):Corpus of Japanese Dialects)*

3.2. wav2vec2.0-based multi-dialect ASR

wav2vec2.0 consists of a feature encoder layer, a Transformer encoder layer, and a quantization module. The raw audio $\mathbf{X}$ is input to the feature encoder layer, and the resulting T-frame feature representation $\mathbf{Z}$ is input to the Transformer encoder layer which outputs a T-frame audio representation $\mathbf{C}$. At this time, some frames of $\mathbf{Z}$ are masked. The quantization module takes $\mathbf{Z}$ as input and outputs a quantized representation $\mathbf{Q}$. The learning objective L_{SSL} optimizes the contrast loss L_m, augmented by the diversity loss L_d. In this study, we compare and investigate the following three fine-tuning methods for multi-dialect ASR with the SSL model of wav2vec2.0.

3.2.1. Multitask learning of ASR and DID

ASR and DID multitask learning is one of the promising ways to improve a multi-dialect ASR [2, 24, 25], as it is expected to improve the performance by simultaneously predicting which local dialect is used. Therefore, we adopt this method in the SSL-based ASR framework as one of the methods to compare. The structure of the ASR&DID joint fine-tuning model is shown in Figure 1. The ASR task computes the CTC loss L_{CTC} from the speech representation $\mathbf{C}$, while the DID task computes the cross-entropy (CE) loss L_{CE} from the frame average of $\mathbf{C}$.

$$L = \alpha L_{CTC} + (1-\alpha) L_{CE} \tag{2}$$

where α is an adjustable parameter. Hereafter, this fine-tuning method is referred to as ft(ASR&DID) for short.

3.2.2. Joint training with SSL and ASR losses

The effectiveness of adapting from the source language to the target language by simultaneously considering SSL loss and downstream task loss has been shown in previous works [17, 18, 19]. Therefore, as a second method, we employ joint training with SSL loss and CTC loss of the ASR task.

$$L = L_{CTC} + \beta L_{SSL} \tag{3}$$

where β is an adjustable parameter. Hereafter, this fine-tuning method is referred to as ft(SSL&ASR) for short.

3.2.3. Joint training with SSL, ASR, and DID losses

We propose joint training with SSL, CTC, and CE losses to learn knowledge of the three objectives at once.

$$L = \alpha L_{CTC} + (1-\alpha) L_{CE} + \gamma L_{SSL} \tag{4}$$

where γ is an adjustable parameter. The flow of adaptation from a standard language ASR model to a multi-dialect model using this method is shown in Figure 2. In the first stage, the SSL and ASR losses are jointly optimized for standard language corpus; in the second stage, the SSL, ASR, and DID losses are jointly optimized for dialect corpus; and in the third stage, only ASR and DID losses are jointly optimized for dialect corpus. We denote this model as " XLSR+ft(SSL&ASR)$_S$+ft(SSL&ASR&DID)$_D$+ft(ASR&DID)$_D$ " and similar abbreviations are used in the next section, where ft stands for fine-tuning and the subscripts S and D stand for the standard and dialect corpus used for fine-tuning, respectively.

4. Experiments

4.1. Corpus

The Corpus of Japanese Dialects (COJADS) [3] is used as the target multi-dialect corpus (denoted as D). It is an extremely realistic dialect corpus that contains natural dialects which tend to appear in spoken language all over Japan since it contains a large number of multi-person discourse speech with low recording quality. It occasionally includes utterances with low transcription accuracy. The transcription is available in katakana only. We used 65h of training data and 2h of evaluation data. The evaluation data are selected so that there is no overlap between the training data and the speakers, also the distribution of age, gender, and region are considered to be as close as possible. For the validation data, 4k utterances are randomly selected from the training data. The Corpus of Spontaneous Japanese (CSJ) [26] is used as the standard language corpus (denoted as S). It was recorded in a clean environment and the transcriptions were maintained. Only katakana is used in the experiment. 230h of monologue speech lectures are used as training data and 6h as validation data. The eval1 test dataset of CSJ is used for evaluation.

4.2. Implementation Details

In the experiments related to the XLSR, fairseq toolkit [27] is used. We use pretrained XLSRs with 53 languages and 56 kh of speech. Japanese is included in the 53 languages, but only in a very small amount (2h). A LARGE model with 7 layers of feature encoders and 24 layers of Transformer encoders is used. The maximum number of samples in a batch is set to 1.28m, or 1.2m if SSL learning is included. The learning rate is 3×10^{-5} and a tri-state larning rate schedule [6] is used. The number of updates is set to 25k∼35k; learning involving ASR&DID

Table 1: *Results of ASR model with/without SSL for multi-dialect speech evaluation dataset (COJADS). Subscripts S and D stand for the standard and multi-dialect corpus used for fine-tuning respectively.*

Model	CER(%)	RTF
Hybrid-CTC/Transformer$_\mathrm{D}$	47.0	0.131
XLSR+ft(ASR)$_\mathrm{D}$	**40.6**	0.003
XLSR+ft(ASR)$_\mathrm{S}$	60.3	0.003

Table 2: *Results of ASR model with/without SSL for standard Japanese evaluation dataset (CSJ eval1)*

Model	CER(%)	RTF
Hybrid-CTC/Transformer$_\mathrm{S}$	5.2	0.188
XLSR+ft(ASR)$_\mathrm{S}$	**4.2**	0.003
XLSR+ft(ASR)$_\mathrm{D}$	16.1	0.003

multitask tended to slow convergence. When SSL learning is not included, no masking process is performed. Single GPU is used. Other settings are basically the defaults of fairseq. Examinations revealed that the appropriate values for α are 0.1 for ASR and 0.01 for DID tasks. The values of β and γ are both found to be 0.5, which is appropriate for dialect corpus, so these values are used. For the Hybrid-CTC/Transformer ASR, experiments are conducted on Espnet2 toolkit [28]. The model structure is 12-layer encoder and 6-layer decoder. The learning rate is 5×10^{-3}. The values of λ is 0.3. The number of epochs is set to 50. The beam size during inference is set to 10.

4.3. Comparison of SSL & non-SSL ASR models

The Hybrid-CTC/Transformer model is compared to the fine-tuned XLSR ASR. The results of the comparison of character error rates (CER) for COJADS and CSJ evaluation data are shown in Table 1 and 2. The fine-tuned XLSR ASRs perform better for both the COJADS and CSJ corpus. This indicates that the knowledge obtained from pretraining works effectively for ASR. The very high CER of COJADS compared to CSJ indicates that dialect speech recognition for diverse topics and conversational speech in a real environment is a challenging task. In addition, analysis of utterances contained more than 50% deletion or insertion errors in the evaluation data revealed that about 10% or more of the utterances are difficult to predict, including those with poor transcription accuracy and those spoken by multiple people at the same time. The problem is that these utterances reduce the overall performance. In terms of inference time using a GPU, the Real Time Factor (RTF) is smaller for models with XLSR because Hybrid-CTC/Transformer uses beam search to predict in an autoregressive manner.

4.4. Effect of additional SSL fine-tuning for each corpus

A model adapted with SSL fine-tuning from XLSR using the respective corpus of COJADS and CSJ is created, and the results of ASR fine-tuning from there for COJADS are shown in Table 3. By performing SSL fine-tuning on the COJADS, in addition to recording environment and age, Japanese dialectal knowledge is implicitly learned, resulting in a speech representation more suitable for the ASR of the COJADS. Even when SSL fine-tuning is performed in the CSJ, an adaptive effect on the ASR of the COJADS is observed by learning knowledge of standard Japanese.

Table 3: *Results of additional SSL fine-tuning of XLSR with standard (S) / multi-dialect (D) speech datasets*

Model	CER(%)
XLSR+ft(ASR)$_\mathrm{D}$	40.6
XLSR+ft(SSL)$_\mathrm{D}$+ft(ASR)$_\mathrm{D}$	**38.8**
XLSR+ft(SSL)$_\mathrm{S}$+ft(ASR)$_\mathrm{D}$	39.2

Table 4: *Results of multitask learning of XLSR with ASR and DID tasks*

Model	α	CER(%)	Acc(%)
XLSR+ft(SSL)$_\mathrm{D}$			
+ft(ASR)$_\mathrm{D}$	1.0	38.8	-
+ft(DID-8)$_\mathrm{D}$	0	-	91.9
+ft(ASR&DID-8)$_\mathrm{D}$	0.1	38.6	90.1
+ft(ASR&DID-8)$_\mathrm{D}$	0.01	40.0	**92.6**
+ft(ASR&DID-17)$_\mathrm{D}$	0.1	**38.1**	90.4

4.5. Effect of ASR&DID multitasking learning

Table 4 shows the results comparing models fine-tuned for three tasks ASR, DID, and joint ASR&DID from the SSL fine-tuned model of XLSR using COJADS. Two types of DID are compared: eight local classifications and 17 local classifications that are further subdivided from 8 local classes. The DID fine-tuned model shows high classification accuracy. The result suggests that multilingual pretraining worked effectively for DID task as well as for the language identification task in [9]. However, when DID is performed on non-speech portions[1] of 0.5 seconds each from all recordings, an accuracy of about 65% is obtained, indicating that differences in recording conditions from region to region are also learned. Therefore, we think it is necessary to consider DID modeling that is robust to various recording environments as a future challenge. On the other hand, by adjusting the α of the ASR&DID joint fine-tuning model, higher performance is obtained than the respective singletask models. This indicates that through joint fine-tuning, ASR improves performance by incorporating DID's knowledge and DID improves performance by incorporating ASR's knowledge. In addition, the effect of joint fine-tuning on ASR is enhanced when DID is subdivided into 17 regions instead of 8.

4.6. Effect of SSL joint fine-tuning from XLSR

Table 5 shows the results comparing the XLSR to SSL&ASR joint fine-tuning and SSL&ASR&DID joint fine-tuning models using COJADS. Compared to the SSL fine-tuning model, the performance of the SSL&ASR joint fine-tuning is slightly lower and the SSL&ASR&DID joint fine-tuning is slightly higher. It is believed that learning the knowledge of ASR and DID along with SSL allows efficient ASR&DID joint fine-tuning.

4.7. Effect of incorporating standard language corpus and joint fine-tuning

An SSL&ASR jointly fine-tuned model is created from the XLSR using CSJ. We compare the SSL&ASR jointly fine-tuned

[1] As listening to the recordings revealed that even label-based non-speech segments often contain speech, we further exclude likely speech segments from them by referring to a range of energy values of 0-4kHz.

Table 5: *Effect of SSL joint training and multitask learning of XLSR-based multi-dialect ASR model. DID-17 task is used for multitask learning.*

Model	α	**CER**(%)	**Acc**(%)
XLSR			
+ft(SSL)$_\text{D}$+ft(ASR)$_\text{D}$	1.0	38.8	-
+ft(SSL)$_\text{D}$+ft(ASR&DID)$_\text{D}$	0.1	38.1	**90.4**
+ft(SSL&ASR)$_\text{D}$+ft(ASR)$_\text{D}$	1.0	39.0	-
+ft(SSL&ASR)$_\text{D}$+ft(ASR&DID)$_\text{D}$	0.1	38.2	88.9
+ft(SSL&ASR&DID)$_\text{D}$+ft(ASR&DID)$_\text{D}$	0.1	**37.9**	90.1

Table 6: *Effect of incorporating standard corpus: SSL&ASR fine-tuning of XLSR with standard speech (CSJ) → two-step fine-tuning with multi-dialect speech (COJADS). DID-17 task is used for multitask learning.*

Model	α	**CER**(%)	**Acc**(%)
XLSR+ft(SSL&ASR)$_\text{S}$			
+ft(SSL&ASR)$_\text{D}$+ft(ASR)$_\text{D}$	1.0	38.5	-
+ft(SSL&ASR)$_\text{D}$+ft(ASR&DID)$_\text{D}$	0.1	37.8	89.0
+ft(SSL&ASR&DID)$_\text{D}$+ft(ASR&DID)$_\text{D}$	0.1	**37.0**	**91.5**

model using COJADS with the SSL&ASR&DID jointly fine-tuned model, and the results are shown in Table 6. Compared to the results in Table 5, the overall ASR performance is improved by learning knowledge of the standard language CSJ and adapting it to the dialect of COJADS. Note also that there is a large performance difference between SSL&ASR joint fine-tuning and SSL&ASR&DID joint fine-tuning. When the final ASRDID joint fine-tuning is performed, the SSL&ASR joint fine-tuning gives a CER of 37.8% and a classification accuracy of 89.0%, while the SSL&ASR&DID joint fine-tuning has a CER of 37.0% and a classification accuracy of 91.5%. The significant increase in classification accuracy from 90.4% in Table 5 indicates that the use of standard language knowledge through SSL&ASR&DID joint fine-tuning improves the performance of DID. The combined effect of the two tasks also increases the performance of ASR task, and the best results for both tasks are obtained in this study. The reason why the joint fine-tuning of SSL&ASR is less effective in adapting to dialects may be because the model loses the knowledge of the standard language required for DID and was unable to fully utilize them.

4.8. Breakdown and analysis of results against COJADS evaluation data

The distribution of CER and classification accuracy per speaker for the fine-tuned (ASR,DID-8) model from XLSR using COJADS for the evaluation data is shown in Figure 3. Bars represent speakers, colors represent regions, and lined sections represent prefectures. For ASR, CER increases with transcription accuracy and other factors, regardless of region. On the other hand, for DID, there is a decrease in classification accuracy for dialects that differ significantly from standard language, such as Aomori (2) and Kagoshima (46). The audio of Chiba (12), which shows poor accuracy, contained the chirping of crickets, indicating that in special recording environments, the classification is more acoustic than dialectal and accuracy is reduced.

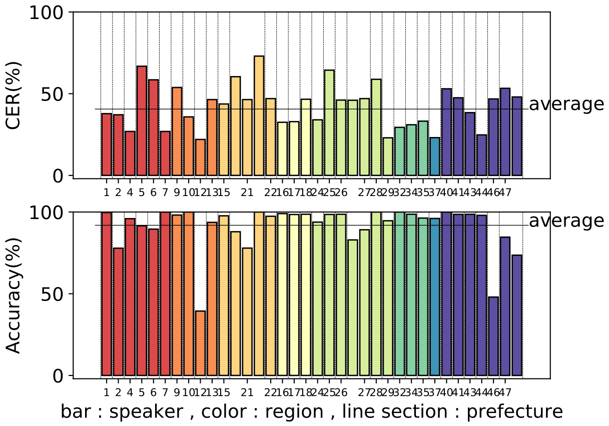

Figure 3: *Performance distribution of (ASR,DID-8) fine-tuned model for COJADS from XLSR by speaker for COJADS evaluation data. Top: ASR, Bottom: DID*

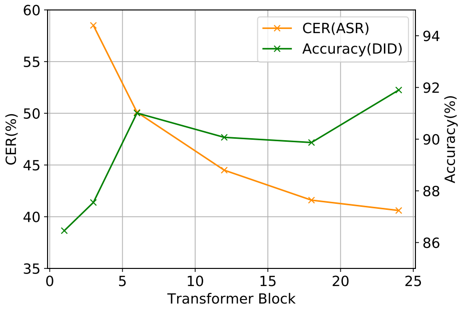

Figure 4: *Results of COJADS evaluation for models fine-tuned (ASR, DID-8) to COJADS using XLSR intermediate output*

4.9. Fine-tuning using XLSR intermediate output

It is considered that the XLSR contains different knowledge in each Transformer layer [16]. Figure 4 shows the layer-by-layer results of the (ASR,DID-8) fine-tuned model for COJADS. For example, the results at layer 3 are the result of fine-tuning by clipping up to 3 layers of the XLSR. DID shows a classification accuracy of more than 90% up to layer 6 and no change up to layer 18. This indicates that knowledge from layer 6 to 18 may be unnecessary for the DID task. The use of knowledge in the latter 18∼24 layers further increases the accuracy of the DID task. This may be due to the fact that linguistic knowledge near the last layer is effective for DID. On the other hand, knowledge in all layers is effective for ASR, and performance increases with each layer.

5. Conclusion

This study presents a baseline for dialect speech recognition of Japanese using a publicly available Corpus of Japanese Dialects. We used the XLSR self-supervised learning model and adapted it from a standard language corpus to a dialect corpus by combining multitask learning on SSL, ASR, and DID. The proposed method achieves a relative character error rate reduction of 8.9% from a simple ASR fine-tuned model. Plan for the future work includes developing an algorithm that is robust to the inaccurate transcription of Corpus of Japanese Dialects and to the recording environment.

6. References

[1] R. Imaizumi, R. Masumura, S. Shiota, and H. Kiya, "Dialect-Aware Modeling for End-to-End Japanese Dialect Speech Recognition," in *Proc. 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Incheon, Korea, Dec. 2020, pp. 297–301.

[2] R. Imaizumi, R. Masumura, S. Shiota, and H. Kiya, "End-to-end Japanese Multi-dialect Speech Recognition and Dialect Identification with Multi-task Learning," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, Mar. 2022.

[3] "COJADS." [Online]. Available: https://www2.ninjal.ac.jp/cojads/index.html

[4] Y. Chung, W. Hsu, H. Tang, and J. Glass, "An Unsupervised Autoregressive Model for Speech Representation Learning," in *Proc. INTERSPEECH 2019 – Annual Conference of the International Speech Communication Association*, Graz, Austria, Sep. 2019, pp. 146–150.

[5] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-Training for Speech Recognition," in *Proc. INTERSPEECH 2019 – Annual Conference of the International Speech Communication Association*, Graz, Austria, Sep. 2019, pp. 3465–3469.

[6] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[7] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[8] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, and F. Wei, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1–14, 2022.

[9] A. Tjandra, D. G. Choudhury, F. Zhang, K. Singh, A. Conneau, A. Baevski, A. Sela, Y. Saraf, and M. Auli, "Improved Language Identification Through Cross-Lingual Self-Supervised Learning," in *Proc. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May. 2022, pp. 6877–6881.

[10] J. Shor, A. Jansen, W. Han, D. Park, and Y. Zhang, "Universal Paralinguistic Speech Representations Using self-Supervised Conformers," in *Proc. 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May. 2022, pp. 3169–3173.

[11] S. Chen, Y. Wu, C. Wang, S. Liu, Z. Chen, P. Wang, G. Liu, J. Li, J. Wu, X. Yu, and F. Wei, "Why does Self-Supervised Learning for Speech Recognition Benefit Speaker Recognition?" in *Proc. INTERSPEECH 2022 – Annual Conference of the International Speech Communication Association*, Incheon, Korea, Sep. 2022, pp. 3699–3703.

[12] T. Tanaka, R. Masumura, H. Sato, M. Ihori, K. Matsuura, T. Ashihara, and T. Moriya, "Domain Adversarial Self-Supervised Speech Representation Learning for Improving Unknown Domain Downstream Tasks," in *Proc. INTERSPEECH 2022 – Annual Conference of the International Speech Communication Association*, Incheon, Korea, Sep. 2022, pp. 1066–1070.

[13] S. Hussain, V. Nguyen, S. Zhang, and E. Visser, "Multi-Task Voice Activated Framework Using Self-Supervised Learning," in *Proc. 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May. 2022, pp. 6137–6141.

[14] G. Ramesh, C. S. Kumar, and K. S. R. Murty, "Self-supervised Phonotactic Representations for Language Identification," in *Proc. INTERSPEECH 2021 – Annual Conference of the International Speech Communication Association*, Brno, Czechia, Sep. 2021, pp. 1514–1518.

[15] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proc. INTERSPEECH 2021 – Annual Conference of the International Speech Communication Association*, Brno, Czechia, Sep. 2021, pp. 2426–2430.

[16] Jing Zhao and Wei-Qiang Zhang, "Improving Automatic Speech Recognition Performance for Low-Resource Languages With Self-Supervised Models," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1227–1241, Oct. 2022.

[17] C. Wang, Y. Wu, Y. Qian, K. Kumatani, S. Liu, F. Wei, M. Zeng, and X. Huang, "UniSpeech: Unified Speech Representation Learning with Labeled and Unlabeled Data," in *Proc. 2021 International Conference on Machine Learning*. PMLR, Jul. 2021, pp. 10 937–10 947.

[18] S. Raghavan and K. Shubham, "Hybrid unsupervised and supervised multitask learning for speech recognition in low resource languages," in *Proc. Workshop on Machine Learning in Speech and Language Processing*, 2021.

[19] J. Bai, B. Li, Y. Zhang, A. Bapna, N. Siddhartha, K. Chai Sim, and T. N. Sainath, "Joint Unsupervised and Supervised Training for Multilingual ASR," in *Proc. 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May. 2022, pp. 6402–6406.

[20] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent NN: First results," in *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[21] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/Attention Architecture for End-to-End Speech Recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, pp. 1240–1253, Dec. 2017.

[22] S. Karita, N. E. Y. Soplin, S. Watanabe, M. Delcroix, A. Ogawa, and T. Nakatani, "Improving Transformer-based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration," in *Proc. INTERSPEECH 2019 – Annual Conference of the International Speech Communication Association*, Graz, Austria, Sep. 2019, pp. 1408–1412.

[23] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. international conference on Machine learning*, 2006, pp. 369–376.

[24] A. Yadavalli, G. Mirishkar, and A. K. Vuppala, "Multi-Task End-to-End Model for Telugu Dialect and Speech Recognition," in *Proc. INTERSPEECH 2022 – Annual Conference of the International Speech Communication Association*, Incheon, Korea, Sep. 2022, pp. 1387–1391.

[25] Y. Zhao, J. Yue, X. Xu, L. Wu, and X. Li, "End-to-end-based Tibetan Multitask Speech Recognition," *Transaction on Access*, vol. 7, pp. 162 519–162 529, Nov. 2019.

[26] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous Speech Corpus of Japanese," in *Proc. International Conference on Language Resources and Evaluation (LREC'00), Athens, Greece. European Language Resources Association (ELRA)*, 2000.

[27] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A Fast, Extensible Toolkit for Sequence Modeling," in *Proc. NAACL-HLT 2019: Demonstrations*, 2019.

[28] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-End Speech Processing Toolkit," in *Proc. INTERSPEECH 2018 – Annual Conference of the International Speech Communication Association*, Hyderabad, India, Sep. 2018, pp. 2207–2211.