Domain Adaptation by Alternating Learning of Acoustic and Linguistic Information for Japanese Deaf and Hard-of-Hearing People

TAKAHASHI, Kaito*, WAKABAYASHI, Yukoh*, OHTA, Kengo†, KOBAYASHI, Akio‡ and KITAOKA, Norihide*
* Toyohashi University of Technology, Japan
† National Institute of Technology, Anan College, Japan
‡ Yamato University, Japan

***Abstract*—More than half of Japanese people with hearing impairments communicate using speech, however speech recognition systems trained using speech from individuals with normal hearing are unable to achieve sufficient recognition accuracy of speech from individuals with hearing impairments. Therefore, speech recognition systems adapted for individuals with hearing impairments are needed. In this study, we propose a learning method that retains both acoustic and linguistic information to achieve more accurate recognition of speech of the hearing-impaired. Our proposed method performs domain adaptation by alternately switching whether to train the Transformer encoder layer and decoder based on the input speech. By using this method, we can create a speech recognizer adapted to hearing-impaired speech acoustically, while preserving the linguistic information of a general training corpus, thereby improving recognition performance for hearing-impaired speech.**

I. Introduction

In recent years, the accuracy of automatic speech recognition (ASR) has improved, and it is now being utilized in various scenarios, such as smart speakers, voice assistants, and voice input. The ASR systems used in these applications are generally trained with the speech of hearing individuals, and can achieve high recognition accuracy for standard speech. However, it has been reported that such models have low recognition accuracy for the speech of hearing-impaired individuals [1]. Approximately 25% of hearing-impaired individuals are said to use sign language as a means of communication in their daily lives [2], but smooth communication through sign language requires both parties to understand it. Additionally, it has been reported that more than half of hearing-impaired individuals use speech to communicate, but their speech tends to be difficult to understand, thus the use of high-accuracy ASR would be very helpful. However, existing speech recognizers are unable to achieve sufficient recognition performance.

One obstacle to achieving sufficient recognition accuracy of the speech of the hearing-impaired is the lack of a corpus of their speech data. Another is the fact that their speech has different acoustic characteristics than that of hearing individuals, in terms of articulation, prosody, and phonation, which are factors that reduce speech recognition accuracy. Research has been conducted on speech recognition of the speech of individuals with articulation disorders, which has characteristics similar to the speech of hearing-impaired individuals [3], [4]. To overcome the problem of a lack of speech data for individuals with articulation disorders, methods have been proposed to adapt standard speech recognition models to recognize such speech [5], [6]. Furthermore, methods using self-supervised models pre-trained with a large amount of unlabeled data have also been proposed [7]. Self-supervised learning (SSL) has achieved high accuracy in various tasks such as speech recognition [8], [9], speech emotion recognition [10], and speaker identification [11]. Moreover, it has been reported that speech representations generated using SSL-based speech recognition models are robust to domain mismatches [7], [12], [13]. Pasad et al. [14] have shown that the layer-wise representations of wav2vec 2.0 [8], a framework that learns speech representations from audio only, follow an acoustic-linguistic hierarchy. They have also shown that the weights of the upper layers of a pre-trained wav2vec 2.0 framework are not suitable for ASR fine-tuning and ASR fine-tuning using wav2vec 2.0 can be improved by initializing the weights of the upper layers. It has also been reported that speech features captured by wav2vec 2.0 representation, particularly the speech features of XLSR-53 [15], are effective for improving recognition of the speech of individuals with articulation disorders [7]. However, the methods proposed in these studies are mostly for acoustic domain adaptation, and linguistic information also needs to be considered to achieve high speech recognition accuracy. John et al. [16] proposed constructing an ASR model which can recognize out-of-vocabulary words in the speech of individuals with articulation disorders by converting normal speech containing unknown words into the speech of individuals with articulation disorders, and using the converted speech as training data. However, this method relies on the accuracy of the voice conversion used for acoustic domain adaptation.

Therefore, in this study we aim to construct an ASR that retains acoustic speech representations obtained through SSL, in addition to the linguistic information obtained from a general, large-scale speech corpus. Motivated by our previous research [17], which is currently under review, we first consid-

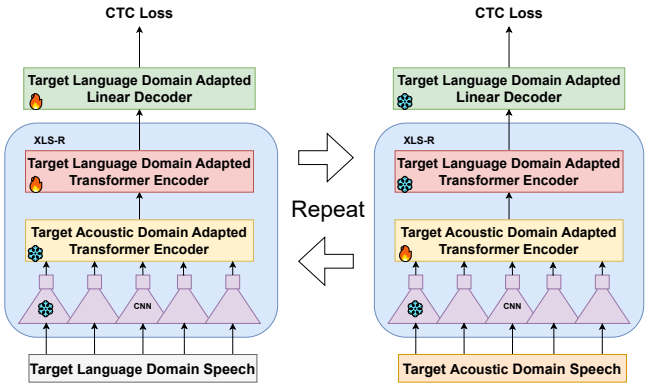

Fig. 1. Proposed alternating learning of acoustic and linguistic information.

ered replacing the encoder layers, but observed overfitting after fine-tuning the ASR with a small amount of data after layer replacement. To further improve the accuracy of recognition of speech from hearing-impaired individuals, we propose a learning method that suppresses overfitting while retaining the linguistic information obtained from a general, large-scale training corpus, and which also adapts the model to hearing-impaired speech using the available target domain speech. This process of acoustic and linguistic adaptation are repeatedly performed during model training. The contributions of this paper are as follows:

- We demonstrate that the proposed method constructs a speech recognition model that retains both acoustic information from hearing-impaired speech and linguistic information from of standard speech.
- We show that a model trained using the proposed method outperforms the same model fine-tuned in a conventional manner, as well as other models with a larger number of parameters, in terms of recognition accuracy for hearing-impaired speech.

II. Proposed Method

Our previous work [17], which is currently under review, proposed a method of replacing the encoder layers of an ASR model in order to adapt it to the processing hearing-impaired speech, using the acoustic information of speech from hearing-impaired individuals while retaining the linguistic information obtained from a general, large-scale speech corpus. In this method, fine-tuning the model with the speech of hearing-impaired individuals after replacing the encoder layers. However, we observed that fine-tuning with a small amount of speech data from hearing-impaired individuals caused overfitting of the linguistic information. To prevent this overfitting, in this paper we propose a learning method that alternately learns the acoustic information of hearing-impaired speech and the linguistic information of a general, large-scale speech corpus, after replacing the encoder layers. This method is inspired by the observation that the lower layers in the Transformer encoder of the ASR model process acoustic information while the upper layers process linguistic information. The proposed method uses as its initial state the model with replaced encoder layers described in our previous study, currently under review [17], as described in Section II-B.

A. Alternating learning

A diagram of the proposed method is shown in Fig. 1. In the proposed method, Connectionist Temporal Classification (CTC) [18] fine-tuning is performed as follows:

1) When using the speech of hearing individuals as the target language domain, the lower layers of the Transformer encoder are frozen, and only the upper layers close to the output layer, and the decoder, are trained.
2) When using the speech of hearing-impaired individuals as the target acoustic domain, the upper layers of the Transformer encoder are frozen, and only the lower layers close to the input layer are trained.
3) Repeat steps 1) and 2).

Using this training method, the ASR can be domain-adapted to both the target language domain and the target acoustic domain.

B. Initialization of ASR model

We initialize the ASR model using the following procedure prior to the alternating learning, described in Section II-A. The model is initialized using the method shown in Fig. 2. This approach is based on a method proposed in our previous paper [17], which is currently under review, involving the following three steps:

1) Additional pretraining: We further pre-train an XLS-R model to adapt it to the the acoustic information of speech from hearing impaired individuals, using both large scale, standard Japanese speech data, and the speech data of hearing-impaired individuals, both of which will also be used during subsequent fine-tuning.

2) 1st fine-tuning: To learn linguistic information, we perform the first fine-tuning of the speech recognition model using a large amount of speech data from hearing individuals the target linguistic domain. We also add a single, fully-connected layer as the decoder and freeze the CNN encoder.

3) Replace layers: Assuming that the effect of acoustic domain adaptation from the additional pre-training has been forgotten due to the first fine-tuning, we construct a speech recognition model that retains both acoustic and linguistic information by replacing part of the encoder layers of the fine-tuned speech recognition model with the pre-trained XLS-R model's encoder layers. In this study, the initial state was set by replacing the lower half of the Transformer encoder layers.

III. Experimental Setup

A. Hearing-impaired speech corpus

The corpus of speech from hearing-impaired individuals which was used in this study was recorded by Kobayashi et al. [1], and also includes additional speech data recorded subsequently by the same researchers. This corpus includes

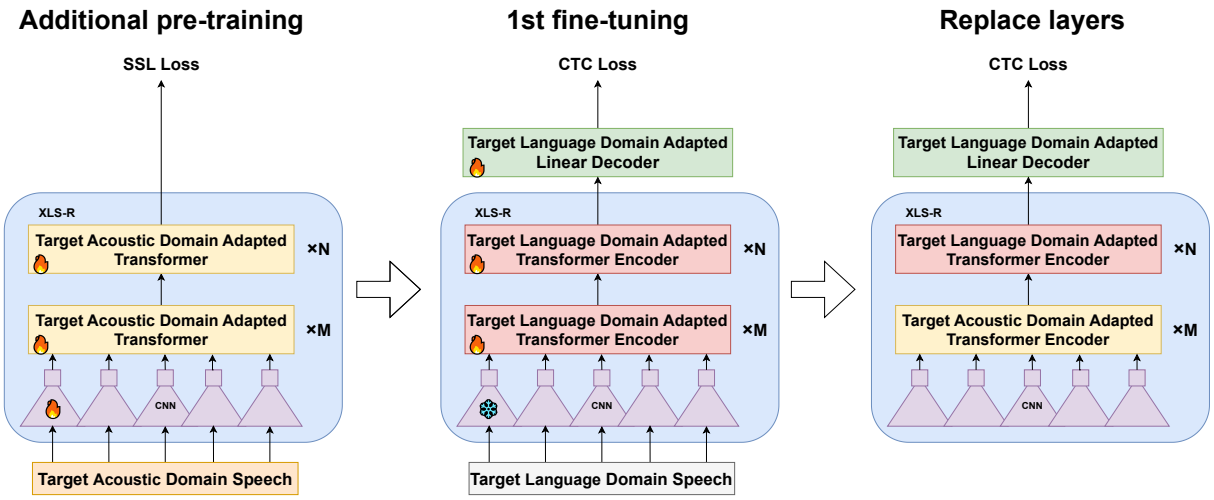

Fig. 2. Initialization of the ASR model before alternating learning. This process includes additional pre-training using SSL and Transformer, fine-tuning of the ASR model by combining the encoder and CTC decoder, and replacing some of the encoder layers.

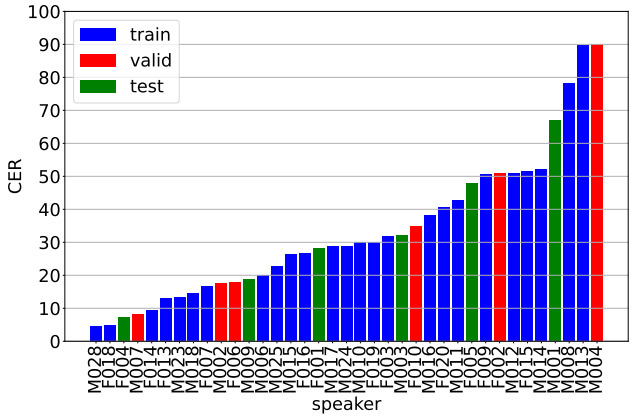

Fig. 3. Division at a splitting of DEAF corpus data into training, validation and testing sets, based on recognition results for each hearing-impaired speaker when using in an ASR model trained with standard on hearing individuals' speech.

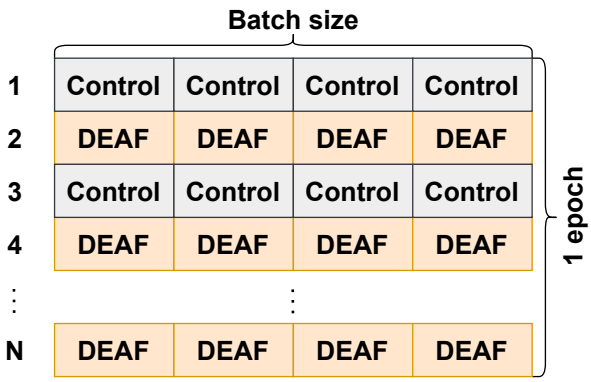

Fig. 4. Order of the training datasets during one epoch of alternating learning, where $1, 2, 3, 4, \ldots, N$ represent the iterations, "Control" and "DEAF" indicate data sampled from the corpus of standard Japanese speech and the corpus of hearing-impaired speech, respectively.

parts of the ATR phoneme-balanced 503 sentences used for the JNAS read speech corpus [19], as read by hearing-impaired individuals. Not all of the speakers read all of the selected ATR sentences, however all of the speakers read both the B and C sets of sentences. Figure 3 shows the Character Error Rate (CER) for the speech recognition of each hearing-impaired participant, sorted in ascending order, when using an ASR model trained only using speech from non-impaired individuals. Based on these results, the corpus of hearing-impaired speech was proportionally divided according to the recognition difficulty of the data, into training, validation and test sets. During this division, care was taken to ensure there was no overlap between the speakers and the utterance content, and differences between the hearing-impaired speech and speech from hearing individuals were considered. In Fig. 3, the colors of the bars represent which of the datasets (training, validation or testing) a speaker's speech was assigned to. The validation set consists of the ATR 503 B set, the test set consists of the ATR 503 C set, and the training set consists of the remaining sets. The training set comprises approximately 16 hours of speech from 16 individuals (6 females and 10 males), while the validation and evaluation sets each comprise approximately 30 minutes of speech from 3 females and 3 males. In this study, we refer to this corpus as the DEAF corpus.

B. Control speech corpus

We used speech from the JNAS corpus of read Japanese as our corpus of standard speech of hearing individuals because the DEAF corpus also consists of readings of the ATR phoneme-balanced 503 sentences included in the JNAS

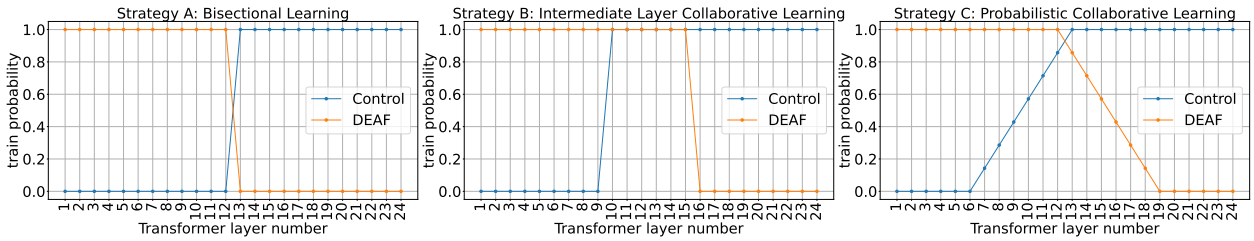

Fig. 5. The probability of learning for each layer of the Transformer encoder

corpus. Therefore, as for the DEAF corpus, the validation set for the corpus of hearing individuals' speech consisted of the ATR 503 B set, the test set consisted of the ATR 503 C set, and the training set consisted of the remaining speech. As a result, the training set comprised approximately 80 hours of speech, the validation set approximately 2 hours, and the evaluation set approximately 3 hours. In addition, the large-scale, Japanese-language, Laboro TV Speech (LTV) corpus [20], which includes approximately 2,000 hours of general television speech, and 767 hours of news and report reading speech, was added to the standard/hearing speaker training set.

C. ASR Model

The XLS-R (0.3B) [21] model, which is based on SSL and includes a Transformer encoder with 24 layers, was used in this study as the ASR encoder. The XLS-R model is often used for downstream tasks after fine-tuning, including speech recognition. Performing SSL using on large-scale speech data allows high speech recognition accuracy, even when the amount of labeled data available for fine-tuning is limited. It has been reported that high recognition accuracy can also be achieved for dysarthric speech, a type of disordered speech, through the use of SSL representations [7]. During fine-tuning, a single full-connected layer was added as the decoder, and training was performed using CTC loss [18].

We set the baseline for ASR performance as the case where fine-tuning is performed on the DEAF corpus after initialization by layer replacement [17]. ASR accuracy when using Whisper Medium [22], which has approximately twice the parameters of XLS-R (0.3B), and ReazonSpeech v2.0 [23] were used for comparison. Whisper Medium is a model that, like XLS-R (0.3B), has a 24-layer Transformer encoder. ReazonSpeech v2.0 is an ASR model trained with a large-scale, Japanese speech dataset that achieves high recognition accuracy for standard speech. Each model was acoustically domain-adapted by fine-tuning only the encoder, using the DEAF corpus.

D. Preparation of the dataset

Our proposed method uses speech from both hearing-impaired and unimpaired individuals, alternately, during model training, which requires equalizing the number of utterances from each corpus which are input during each epoch. However, the DEAF corpus has significantly fewer utterances than the Control corpus. To remove this bias, we first aligned the number of utterances in the DEAF and Control corpora during each epoch by copying the DEAF corpus. Next, we established a data sampling method to alternately input utterances from the Control corpus and the DEAF corpus into the model. The training dataset created from the Control and DEAF corpora is shown in Fig. 4. In each data batch, audio and its corresponding transcription text are sampled from the same corpus, and in the following batch, the same types of data is sampled from the other corpus. In other words, during each iteration data is sampled from the same corpus and input into the model (from either the DEAF or Control corpus), thereby performing alternating training of the model, i.e., the model is trained through alternative learning.

Algorithm 1 Learning Decision Based on Uniform Random Sampling and Threshold

```
Sample p ~ U(0, 1)
if p ≤ threshold then
    Train : Unfreeze layer
else
    Skip : Freeze layer
end if
```

E. Training Strategy

In the Transformer encoder used for ASR, it is assumed that the lower layers process acoustic information, while the upper layers process linguistic information. Some evidences of this has been suggested in studies such as [14], but it is still unclear exactly how each layer processes information. Therefore, in this experiment, we designed three learning strategies and compared the resulting recognition accuracies of the three ASR models. Each of these learning strategies are illustrated in Fig. 5. These learning strategies determine the learning probability, for each layer of the Transformer encoder, for both the Control and DEAF corpora. By comparing this learning probability to a threshold, Algorithm 1 determines probabilistically whether or not to train each layer of the Transformer encoder. The three learning strategies can be described as follow:

- **Strategy A: Bisectional Learning**
  We hypothesized that the lower layers of the Transformer encoder process acoustic information, while the upper

TABLE I
ASR ACCURACY WHEN USING EACH ENCODER TRAINING STRATEGY

Strategy	JNAS CER (%)	DEAF CER (%)
A	8.3	22.9
B	7.9	21.7
C	**7.7**	**21.3**

layers process linguistic information, and divided the layers to be trained with each speech corpus accordingly. When the input is DEAF speech, the upper layers of the Transformer encoder are frozen, and when the input is Control speech, the lower layers are frozen. In other words, the lower 12 layers of the Transformer encoder are trained with the DEAF corpus, and the upper 12 layers are trained with the Control corpus.

- **Strategy B: Intermediate Layer Collaborative Learning**
  This is a modification of Strategy A, in which the intermediate layers of the Transformer are trained with both corpora. This is done because it is unclear whether the intermediate layers of the Transformer encoder process acoustic or linguistic information, and by modifying Strategy A, we can add a margin to the layers trained with each corpus. In other words, the lower layers of the Transformer encoder are trained with the DEAF corpus, the intermediate layers are trained with both the Control and DEAF corpora, and the upper layers are trained with the Control corpus.
- **Strategy C: Probabilistic Collaborative Learning**
  This is a modification of Strategy A where the intermediate layers are trained probabilistically. In Strategy B, we hypothesized that among the intermediate layers trained using both corpora, the closer a layer is to the input layer, the more likely it processes acoustic information, and the closer it is to the output layer, the more likely it processes linguistic information. Therefore, in Strategy C we chose to select which intermediate layers were to be trained with each corpus probabilistically, as an alternative to Strategy B.

IV. Experimental Results

A. Comparison of CER among training strategies

Table II shows CER with the three training strategies. Among the three Encoder training strategies, Strategy C achieved the highest recognition accuracy for both JNAS (unimpaired) and DEAF (impaired) speech. This indicates that Strategy C allowed the ASR model to learn the acoustic information from the DEAF corpus and the linguistic information from the control speech corpus more effectively than the other strategies, therefore this training strategy was used in our proposed method. Additionally, the use of Strategy B resulted in higher recognition accuracy than Strategy A for both JNAS and DEAF speech. These results suggest that having the Encoder layers trained with both corpora leads to better learning outcomes.

B. Comparison with other models

A comparison of speech recognition accuracy in terms of CER for both unimpaired and hearing-impaired speech, between the model trained using the proposed method, the same model trained using other methods, and other models with more parameters, is shown in Table II. The model trained using the proposed method achieved the lowest CERs for speech from either the JNAS or the DEAF corpora, compared to the same model trained with other methods and with the other models. Note however that the CER for JNAS speech for the XLS-R model trained with the JNAS and LTV datasets was 7.7%, which is the same as the model trained using the proposed method. This suggests that the proposed method was able to acquire the acoustic information of the DEAF corpus while suppressing forgetting of the linguistic information from the JNAS and LTV corpora during acoustic training. These results demonstrate that a speech recognition model constructed using the proposed method can retain both specialized acoustic information as well as standard linguistic information.

V. Conclusions

In this study, we proposed an alternating training method for ASR models that retains linguistic information obtained from a general, large-scale speech corpus while also being adapting to recognize the acoustic information of hearing-impaired speech, allowing accurate speech recognition of either type of speech. Our experimental evaluation confirmed that our proposed method achieved higher or equivalent speech recognition performance for standard Japanese speech, and higher recognition performance for hearing-impaired speech, than the other methods. Additionally, we tested three Encoder layer training strategies to investigate their effects on speech recognition accuracy, and found that training the intermediate layers probabilistically, using data from one or the other of the two corpora, resulted in the highest ASR accuracy.

In the future, we plan to optimize the layers to be trained, to further improve recognition accuracy.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP23H00995.

References

[1] A. Kobayashi, K. Yasu, H. Nishizaki, and N. Kitaoka, "Corpus Design and Automatic Speech Recognition for Deaf and Hard-of-Hearing People," in *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, 2021, pp. 17–18. DOI: 10.1109/GCCE53005.2021.9621959.

[2] Japanese Ministry of Health, Labour and Welfare, *Survey on Difficulties in Living*. 2018. [Online]. Available: https://www.mhlw.go.jp/toukei/list/dl/seikatsu_chousa_c_h28.pdf.

TABLE II
COMPARISON OF ASR PERFORMANCE WHEN USING EACH METHOD USING CHARACTER ERROR RATE (%), WITH JNAS (STANDARD) AND DEAF (HEARING-IMPAIRED) JAPANESE SPEECH ("$\rightleftarrows$" AND "+" INDICATE ALTERNATING TRAINING AND CONCATENATION OF THE DATASETS, RESPECTIVELY).

method	model	# Parameters (MB)	1st fine-tune	2nd fine-tune	JNAS CER (%)	DEAF CER (%)
Baseline	ReazonSpeech v2.0	619	DEAF	N/A	10.7	26.0
	Whisper Medium	769	DEAF	N/A	9.0	25.1
	XLS-R	319	JNAS + LTV	N/A	**7.7**	39.5
	XLS-R	319	JNAS + LTV	DEAF	8.3	23.0
	XLS-R w/ Replacement [17]	319	JNAS + LTV	DEAF	9.3	22.1
Proposed	XLS-R w/ Alternating learning	319	JNAS + LTV	JNAS + LTV $\rightleftarrows$ DEAF	**7.7**	**21.3**

[3] K. M. Baskar, T. Herzig, D. Nguyen, *et al.*, "Speaker adaptation for Wav2vec2 based dysarthric ASR," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 9, Incheon, KR: International Speech Communication Association, 2022, pp. 3403–3407. DOI: 10.21437/Interspeech.2022-10896. [Online]. Available: https://www.fit.vut.cz/research/publication/12854.

[4] S. Hu, X. Xie, Z. Jin, *et al.*, "Exploring Self-Supervised Pre-Trained ASR Models for Dysarthric and Elderly Speech Recognition," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10097275.

[5] J. Shor, D. Emanuel, O. Lang, *et al.*, "Personalizing ASR for Dysarthric and Accented Speech with Limited Data," in *Interspeech 2019*, ISCA, Sep. 2019. DOI: 10.21437/interspeech.2019-1427. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1427.

[6] R. Takashima, T. Takiguchi, and Y. Ariki, "Two-Step Acoustic Model Adaptation for Dysarthric Speech Recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6104–6108. DOI: 10.1109/ICASSP40776.2020.9053725.

[7] A. Hernandez, P. A. Pérez-Toro, E. Nöth, J. R. Orozco-Arroyave, A. Maier, and S. H. Yang, *"Cross-lingual Self-Supervised Speech Representations for Improved Dysarthric Speech Recognition"*, 2022. arXiv: 2204.01670 [cs.CL].

[8] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, *"wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations"*, 2020. arXiv: 2006.11477 [cs.CL].

[9] S. Chen, C. Wang, Z. Chen, *et al.*, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022, ISSN: 1941-0484. DOI: 10.1109/jstsp.2022.3188113. [Online]. Available: http://dx.doi.org/10.1109/JSTSP.2022.3188113.

[10] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech 2021*, 2021, pp. 3400–3404. DOI: 10.21437/Interspeech.2021-703.

[11] N. Vaessen and D. A. Van Leeuwen, "Fine-Tuning Wav2Vec2 for Speaker Recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7967–7971. DOI: 10.1109/ICASSP43922.2022.9746952.

[12] W.-N. Hsu, A. Sriram, A. Baevski, *et al.*, *"Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training"*, 2021. arXiv: 2104.01027 [cs.SD].

[13] J. Zhao, G. Shi, G.-B. Wang, and W.-Q. Zhang, "Automatic Speech Recognition for Low-Resource Languages: The Thuee Systems for the IARPA Openasr20 Evaluation," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 335–341. DOI: 10.1109/ASRU51503.2021.9688260.

[14] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-Wise Analysis of a Self-Supervised Speech Representation Model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 914–921. DOI: 10.1109/ASRU51503.2021.9688093.

[15] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, *"Unsupervised Cross-lingual Representation Learning for Speech Recognition"*, 2020. arXiv: 2006.13979 [cs.CL].

[16] J. Harvill, D. Issa, M. Hasegawa-Johnson, and C. Yoo, "Synthesis of New Words for Improved Dysarthric Speech Recognition on an Expanded Vocabulary," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6428–6432. DOI: 10.1109/ICASSP39728.2021.9414869.

[17] K. Takahashi, Y. Wakabayashi, K. Ohta, A. Kobayashi, and N. Kitaoka, "Improving Speech Recognition for Japanese Deaf and Hard-of-Hearing People by Replacing Encoder Layers," in *ICAICTA 2024*, 2024, (Under review).

[18] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural 'networks," vol. 2006, Jan. 2006, pp. 369–376. DOI: 10.1145/1143844.1143891.

[19] T. A. S. of Japan, *"ASJ Japanese Newspaper Article Sentences Read Speech Corpus (JNAS)"*.

[20] S. Ando and H. Fujihara, *"Construction of a Large-scale Japanese ASR Corpus on TV Recordings"*, 2021. arXiv: 2103.14736 [cs.SD].

[21] A. Babu, C. Wang, A. Tjandra, *et al.*, *XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale*, 2021. arXiv: 2111.09296 [cs.CL].

[22] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, *"Robust Speech Recognition via Large-Scale Weak Supervision"*, 2022. arXiv: 2212.04356 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2212.04356.

[23] Y. Yin, D. Mori, and S. Fujimoto, "ReazonSpeech: A Free and Massive Corpus for Japanese ASR," *Reazon Holdings, Inc., Clear Code, Inc.*, Feb. 2024. [Online]. Available: https://research.reazon.jp/blog/2024-02-14-ReazonSpeech.html.