EasyChair Preprint

№ 15118

Improving Speech Recognition for Japanese Deaf and Hard-of-Hearing People by Replacing Encoder Layers

Kaito Takahashi, Yukoh Wakabayashi, Kengo Ohta,
Akio Kobayashi and Norihide Kitaoka

September 27, 2024

Improving Speech Recognition for Japanese Deaf and Hard-of-Hearing People by Replacing Encoder Layers

1st TAKAHASHI, Kaito
Toyohashi University of Technology
Aichi, Japan

2nd WAKABAYASHI, Yukoh
Toyohashi University of Technology
Aichi, Japan

3rd OHTA, Kengo
National Institute of Technology, Anan College
Tokushima, Japan

4th KOBAYASHI, Akio
Yamato University
Osaka, Japan

5th KITAOKA, Norihide
Toyohashi University of Technology
Aichi, Japan

***Abstract*—Communication between hearing individuals and those with hearing impairments generally involves sign language, written communication, and speech. It has been reported that more than half of Japanese people with hearing impairments communicate using speech. Therefore, speech recognition systems available for individuals with hearing impairments are demanded. However, speech recognition systems trained on speech from hearing individuals do not achieve high recognition accuracy for speech from individuals with hearing impairments. In this study, we propose a method to replace the encoder layer of the speech recognition model based on SSL to achieve high-accuracy speech recognition for speech from individuals with hearing impairments. By this method, we improved the recognition performance for significantly speech from individuals with hearing impairments.**

***Index Terms*—automatic speech recognition, deaf speech, self-supervised learning, domain adaptation**

I. Introduction

In recent years, the accuracy of speech recognition has improved and is being utilized in various scenarios. For example, smart speakers, voice assistants, and voice input. The speech recognition used in these applications is generally trained on the speech of hearing individuals and achieves high accuracy for their speech. However, it has been reported that models trained on the speech of hearing individuals 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 for smooth communication through sign language, both parties need 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 speech recognition would be helpful. However, at present, existing speech recognizers cannot achieve sufficient recognition performance.

One of the reasons why sufficient recognition accuracy cannot be achieved for speech of hearing-impaired individuals is the lack of speech data from them and the fact that their speech has acoustically different characteristics from that of hearing individuals. The difference is seen in various aspects, such as articulation, prosody, and phonation, which are factors that reduce recognition accuracy in speech recognition. Research on speech recognition is being conducted on speech some of individuals with articulation disorders, which has similar characteristics to the speech of hearing-impaired individuals. To overcome the problem of the lack of data on speech of individuals with articulation disorders, methods are being researched to adapt speech recognition models of healthy individuals to the speech of individuals with articulation disorders [3], [4]. Furthermore, methods using self-supervised models pre-trained on a large amount of unlabeled data are being researched [5]. Self-supervised learning (SSL) has achieved high accuracy in various tasks such as speech recognition [6], [7], speech emotion recognition [8], and speaker identification [9]. Moreover, it has been reported that the speech representations generated by SSL-based speech recognition models are robust to domain mismatches [5], [10], [11]. Pasad et al. [12] have shown that the layer-wise representations of wav2vec 2.0 [6] follow an acoustic-linguistic hierarchy. Furthermore, they have shown that the weights of the upper layers of the pre-trained wav2vec 2.0 are not suitable for ASR fine-tuning and that the performance of ASR fine-tuning can be improved by initializing the weights of the upper layers. It has also been reported that the speech representations of wav2vec 2.0, particularly the speech representations of XLSR-53 [13], are effective for speech recognition in the speech of individuals with articulation disorders [5]. However, these methods are studies on acoustic domain adaptation, and linguistic information also needs to be considered to achieve high recognition accuracy. John et al. [14] proposed to construct on 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

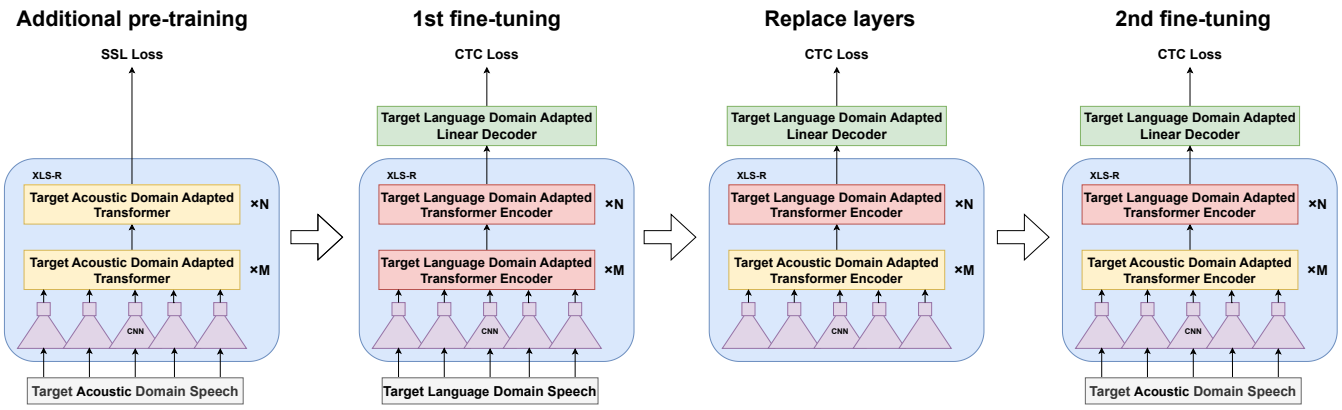

Fig. 1. The flow of constructing the proposed ASR model with gradual fine-tuning and layer replacement.

and using the converted speech as training data. However, this method relies on the accuracy of voice conversion for acoustic domain adaptation.

Therefore, this study aims to combine acoustic speech representations obtained through self-supervised learning using relatively small amount of speech of hearing-impaired individuals with linguistic information obtained from a general large-scale corpus of hearing individuals. To do this, we propose a method to replace some of the encoder layers of the speech recognition model. Our contributions are as follows:

- We demonstrate that additional pre-training can adapt the ASR model to hearing-impaired speech to acoustically.
- We show that by replacing some layers of the encoder of the speech recognition model, it is possible to construct a speech recognition model that retains both the acoustic information of hearing-impaired speech and the linguistic information of general speech.
- The model trained using the proposed method demonstrates superior recognition accuracy compared to the same model simply fine-tuned and other larger models with more parameters.

II. Proposed Method

Figure 1 shows the proposed method. The proposed method consists of the following steps:

A. Additional pretraining

We additional pre-train XLS-R to adapt the acoustic information of speech from hearing-impaired individuals. For additional pre-training, we use Japanese large-scale speech data and speech data of hearing-impaired individuals, which will be used in the subsequent fine-tuning.

B. 1st fine-tuning

To learn linguistic information, we perform the first fine-tuning of the speech recognition model using large-scale hearing individuals speech data from the target linguistic domain. Before the first fine-tuning, we add a single fully-connected layer as the decoder and freeze the CNN encoder.

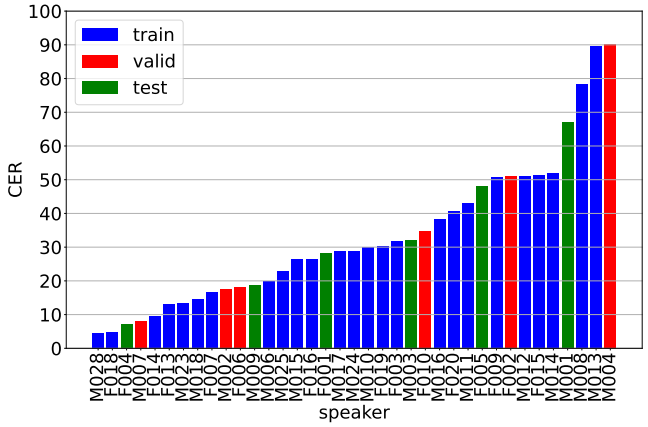

Fig. 2. Data splitting of DEAF corpus based on recognition results for each hearing-impaired speaker in a model trained on hearing individuals' speech

C. Replace layers

We consider that the effect of acoustic domain adaptation from additional pre-training is forgotten due to the first fine-tuning. Therefore, 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 encoder layers.

D. 2nd fine-tuning

Finally, we perform a second fine-tuning of the speech recognition model using speech data from the target acoustic and linguistic domain, which consists of speech from hearing-impaired individuals. Through this process, the speech recognition model becomes adapted to the acoustic information of speech from hearing-impaired individuals and the linguistic information of general speech.

III. Experimental Setup

A. Hearing-impaired speech corpus

The corpus of speech from hearing-impaired individuals consists of the corpus recorded by Kobayashi et al. [1] and

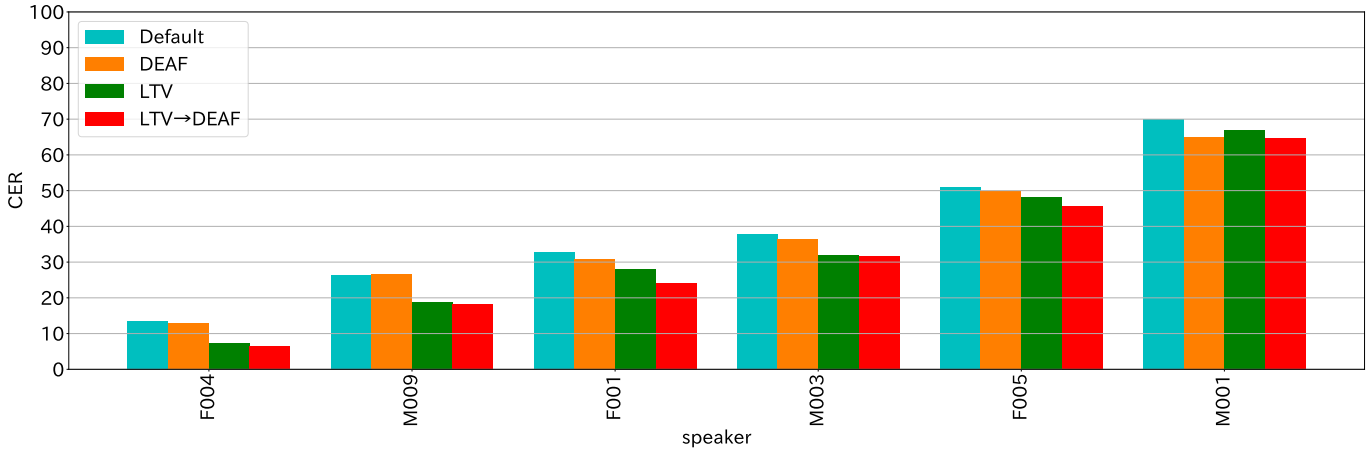

Fig. 3. Evaluation of the Additional pre-training. Default means that there is no additional pre-training.

an additional corpus was recorded subsequently. This corpus includes parts of the ATR phoneme-balanced 503 sentences used for the JNAS read speech corpus [15] read by hearing-impaired individuals. Specifically, not all speakers read all the set sentences; however, at least all speakers read both the B set and the C set. Figure 2 shows the Character Error Rate (CER) for each hearing-impaired speaker, sorted in ascending order, using a model trained only on speech from hearing individuals. Based on these results, the corpus of speech from hearing-impaired individuals was divided for training purposes. During this division, care was taken to ensure that there was no overlap in speakers and utterance content, and the acoustic differences between speech from hearing and hearing-impaired individuals were considered. In Fig. 2, the colors represent the respective divided sets: validation, evaluation, and training sets. The validation set consists of the B set, the evaluation set consists of the 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 the JNAS corpus as the speech corpus for hearing individuals. This is because the DEAF corpus consists of readings of the ATR phoneme-balanced 503 sentences included in JNAS. Therefore, same as the DEAF corpus, the validation set for the hearing individuals' corpus consisted of the B set, the evaluation set consisted of the C set, and the training set consisted of the remaining speech. As a result, the training set comprised approximately 80 hours, the validation set approximately 2 hours, and the evaluation set approximately 3 hours. Additionally, the large-scale Japanese corpus, Laboro TV Speech (LTV) [16], which includes approximately 2000 hours of speech, and 767 hours of news and report readings, was added to the training set.

TABLE I
DEPENDENCE OF CER ON REPLACEMENT OF XLS-R LAYERS

Layer replacement	JNAS CER	DEAF CER
w/o replacement	**8.3**	23.0
1–6	8.4	22.5
1–12	9.3	**22.1**
1–18	11.9	24.9
1–24 (all)	22.8	32.6
7–12	8.8	22.7

C. Model

This study used XLS-R (0.3B) [17] was used as the encoder. This model is based on self-supervised learning and consists of a Transformer Encoder with 24 layers. XLS-R has been applied to various downstream tasks after fine-tuning, including speech recognition, by performing self-supervised learning on large-scale multilingual unlabeled speech data, achieving high accuracy. Through self-supervised learning with large-scale data, high recognition accuracy can be obtained even when the labeled data is limited for fine-tuning to speech recognition. It has been reported that high recognition accuracy can also be achieved for dysarthric speech, a type of disordered speech, using self-supervised learning representations. During fine-tuning, a single fully-connected layer was added as the decoder, and learning was performed using Connectionist Temporal Classification (CTC) loss [18].

Additionally, Whisper Medium [19], which has approximately twice the parameters of XLS-R (0.3B), and ReazonSpeech v2.0 [20] were used for comparison. Whisper Medium is a model that, like XLS-R (0.3B), has a 24-layer Transformer Encoder as its encoder. ReazonSpeech v2.0 is a model trained on a large-scale Japanese speech dataset and achieves high recognition accuracy. Each model was acoustically domain-adapted by fine-tuning only the encoder on the DEAF corpus.

TABLE II
CER OF DEAF, "+" INDICATE THE CONCATENATION OF DATASETS.

method	model	# params (MB)	1st fine-tune	2nd fine-tune	DEAF CER
Baseline	ReazonSpeech v2.0	619	DEAF	N/A	26.0
	Whisper Medium	769	DEAF	N/A	25.1
	XLS-R	319	DEAF	N/A	39.5
	XLS-R	319	JNAS + LTV	DEAF	23.0
Proposed	XLS-R w/ Replacement	319	JNAS + LTV	DEAF	**22.1**

IV. EXPERIMENTAL RESULTS

A. Additional pre-training

To investigate the effectiveness of acoustic domain adaptation through additional pre-training, we compared the CER with and without additional pre-training. The first fine-tuning was performed on JNAS, and the output was in Japanese Katakana characters. The CER for each speaker in the test set of the DEAF corpus, sorted in ascending order, is shown in Fig. 3. First, see the results using LTV to adapt to Japanese. For speakers with low CER without additional pre-training, significant improvement in CER was observed with additional pre-training using LTV. On the other hand, for speakers with high CER without additional pre-training, the effect was limited. This suggests that the speech of speakers with a low CER without additional pre-training had a smaller acoustic gap with hearing individuals' speech, thus the additional pre-training was more pronounced. When additional pre-training was performed on the DEAF corpus, improvement in CER was observed for speakers with a high CER without additional pre-training. Finally, additional pre-training was performed sequentially with LTV and then the DEAF corpus and recognition accuracy improved for all speakers. These results indicate that sequential additional pre-training is effective for various levels of hearing-impaired speech.

B. Layer replacement

It is assumed that increasing the number of replaced Transformer encoder layers can enhance the effect of acoustic domain adaptation. However, increasing the number of replaced layers may reduce the effect of linguistic domain adaptation acquired during the first fine-tuning. To investigate this trade-off, we examined the relationship between the number of replaced layers and recognition accuracy. Table I shows the changes in recognition accuracy when varying the number of replaced layers. In the DEAF corpus, the highest recognition accuracy was achieved when the lower half of the encoder layers were replaced with pre-trained encoder layers. On the other hand, in JNAS, the highest recognition accuracy was achieved when the second fine-tuning was performed without any replacement. Additionally, in JNAS, as the number of replaced layers increased from the lower to the upper layers, recognition accuracy decreased. This is considered to be due to the reduction in the effect of linguistic domain adaptation caused by the replacement. Conversely, in the DEAF corpus, recognition accuracy improved as the number of replaced layers increased up to 12 layers (half of the encoder), but further increases led to a decrease in recognition accuracy. This suggests that while increasing the number of replaced layers enhances the effect of acoustic domain adaptation, it reduces the effect of linguistic domain adaptation.

C. Comparison with other models

The comparison of CER between the model trained using the proposed method and other models with more parameters is shown in Table II. The model trained by the proposed method achieved the lowest CER on the DEAF corpus compared to the same architecture model by conventional method fine-tuned and other models. It was also demonstrated that replacing layers before the second fine-tuning can achieve high recognition accuracy on the DEAF corpus. From the above, it was shown that the proposed method can construct a speech recognition model that maintains effectiveness of both acoustic and linguistic domain adaptation.

V. CONCLUSIONS

In this study, we proposed a method of replacing a part of the encoder layers of a speech recognition model to achieve high-accuracy speech recognition for hearing-impaired speech. Using this method, we confirmed an improvement in recognition performance for hearing-impaired speech. Additionally, we investigated the changes in recognition accuracy with different numbers of replaced layers and found that the highest recognition accuracy was achieved when the lower half of the encoder layers were replaced. In the future, to further improve recognition accuracy, we will explore learning methods that retain both the acoustic and linguistic information of the two corpora. Another future work is searching for the dependence of our method on the domain of data, i.e., whether it is effective for publicly available corpora to confirm its versatility.

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.

[3] 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.

[4] 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.

[5] 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].

[6] 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].

[7] 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.

[8] 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.

[9] 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.

[10] 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].

[11] 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.

[12] 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.

[13] 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].

[14] 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.

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

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

[17] 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].

[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] 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.

[20] 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.