Dysarthric Speech Recognition Using Pseudo-Labeling, Self-Supervised Feature Learning, and a Joint Multi-Task Learning Approach

Takashima, Ryoichi

Sawa, Yuya

Aihara, Ryo

Takiguchi, Tetsuya

Imai, Yoshie

(Citation)
IEEE Access,12:36990-36999

(Issue Date)
2024-03-07

(Resource Type)
journal article

(Version)
Version of Record

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 December 2023, accepted 1 March 2024, date of publication 7 March 2024, date of current version 14 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3374874

RESEARCH ARTICLE

Dysarthric Speech Recognition Using Pseudo-Labeling, Self-Supervised Feature Learning, and a Joint Multi-Task Learning Approach

RYOICHI TAKASHIMA[1], (Member, IEEE), YUYA SAWA[1], RYO AIHARA[2], (Member, IEEE), TETSUYA TAKIGUCHI[1], (Member, IEEE), AND YOSHIE IMAI[2]

[1]Graduate School of System Informatics, Kobe University, Kobe 657-8501, Japan
[2]Information Technology Research and Development Center, Mitsubishi Electric Corporation, Kamakura 247-8501, Japan

Corresponding author: Ryoichi Takashima (rtakashima@port.kobe-u.ac.jp)

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant JP21H00906; and in part by PRESTO, Japan Science and Technology Agency (JST) under Grant JPMJPR23I7.

ABSTRACT In this paper, we investigate the use of the spontaneous speech of dysarthric people for training an automatic speech recognition (ASR) model for them. Although the spontaneous speech of dysarthric people can be collected relatively easily compared to script-reading speech, which is obtained by having them read a prepared script, labeling the spontaneous speech of dysarthric people is very difficult and costly. For training an ASR model using unlabeled speech data, pseudo-labeling and self-supervised feature learning have been studied as effective approaches; however, the effectiveness of these approaches has not been clear when they are applied to the unlabeled dysarthric speech. In addition, pseudo-labeling may not be effective since the pseudo-labels of dysarthric speech include many errors and are not reliable. In this paper, we evaluate the above two approaches for the dysarthric speech recognition, and we propose a multi-task learning approach, which combines these approaches to train an ASR model that is robust against the errors in the pseudo-labels. Experimental results using Japanese and English datasets demonstrated that all approaches are effective, but among them, the proposed multi-task learning approach showed the best performance.

INDEX TERMS Speech recognition, dysarthria, pseudo-labeling, self-supervised feature learning.

I. INTRODUCTION

With the development of deep learning technology, the accuracy of automatic speech recognition (ASR) has been greatly improved and is getting closing to human-level performance [1], [2]. ASR systems have been widely used in applications such as smart speakers, speech translation, car devices, and so on. Because we can input commands to devices with hands-free technology, ASR systems are expected to be applied to input devices for handicapped people, and some applications that make use of ASR systems for handicapped people have already been developed (e.g., Nuance Dragon Anywhere).

Dysarthria is a speech disorder resulting from motor dysfunction, such as cerebral palsy [3] and amyotrophic lateral sclerosis (ALS). Because people suffering from such disease are often restricted their movement of the arms and legs, most of them cannot use alternative means of speech communication, such as sign language or written communication. For this reason, there is a great need of hands-free communication tools and input devices using the ASR system for such people. However, because the speaking style of people having dysarthria tends to be unstable and is greatly different from the speaking style of person without disability, it is difficult for conventional ASR systems to recognize their speech.

Various studies of the dysarthric speech recognition have been conducted. One problem in this task is the scarcity of

The associate editor coordinating the review of this manuscript and approving it for publication was Lin Wang.

the dysarthric speech data for training. Although there have been publicly available databases of dysarthric speech [4], [5], the amount of data is not still sufficient to train an accurate ASR model. For example, speech of eight dysarthric subjects are recorded in TORGO dataset [4]; however, the average length of the recorded speech per subject is less than 30 minutes. In UAspeech dataset [5], speech of fifteen dysarthric subjects are recorded, but only speech of isolated words is recorded. Previous studies tackling this problem of insufficient data have proposed methods including the data augmentation approach [6], [7], [8], [9], [10], the model adaptation approach [11], [12], [13], [14], [15], [16], [17], and an approach that uses multiple datasets of dysarthric speech [18].

One reason why it is difficult to collect speech data from dysarthric people is because they are required to read a prepared script, which is a heavy burden on them, in the aforementioned datasets (speech collected in this way is referred to as "*script-reading speech*"). In this study, we focus on the use of *spontaneous speech* for dysarthric speech recognition. Spontaneous speech of dysarthric people in their daily life can be recorded relatively easily compared to script-reading speech. However, unlike script-reading speech, spontaneous speech requires us to transcribe the speech to obtain the label for training an ASR model, and the transcription of dysarthric speech is extremely difficult and costly due to its being less-intelligible. Therefore, we investigate methods to leverage speech data without labels (i.e., unlabeled data) in this study.

There are two main approaches for utilizing unlabeled speech data for ASR. Although the effectiveness of both two approaches has been confirmed for normal speech recognition, it has not been clear for dysarthric speech recognition. The first is called *pseudo-labeling* [19], [20], [21], which estimates labels from unlabeled speech using speech recognition. The effectiveness of pseudo-labeling depends on the accuracy of the pseudo-label (i.e., performance of the base ASR model). Therefore, pseudo-labeling may not be effective for dysarthric speech recognition since the pseudo-labels of dysarthric speech include many errors.

The second approach is called *self-supervised feature learning* [22], [23], [24], [25], [26], [27], [28], [29], where a neural-network model is trained through a pseudo-task generated without human transcription. The model trained through the self-supervised learning can be used as a good pre-trained ASR model to be fine-tuned using labeled speech data. There have been studies trying to utilize these techniques for dysarthric speech recognition [15], [16], [17]. However, the previous works uses models pre-trained by self-supervised learning using normal speech data, and they use only labeled dysarthric speech data for fine-tuning the pre-trained models. Therefore, the effectiveness of the self-supervised learning with unlabeled dysarthric speech data has not yet been clear.

The contributions of this paper are the following three points. First, we evaluate the use of unlabeled dysarthric speech data with the above two training approaches. Although we conduct the evaluation using TORGO dataset, this dataset contains only a small amount of speech data for each subject, making it difficult to evaluate the effectiveness in detail. For detail evaluation, therefore, we recorded about three hours of unlabeled speech and about one hour of labeled speech from a Japanese dysarthric person. Second, we propose a joint multi-task learning approach combining pseudo-labeling and self-supervised feature learning. While the effectiveness of pseudo-labeling depends on the accuracy of the pseudo-label, self-supervised feature learning is independent of the quality of the pseudo-label because this approach does not use any label information. Therefore, the proposed multi-task learning approach combines self-supervised feature learning with pseudo-labeling to train an ASR model robustly against the errors contained in pseudo-labels. Third, we further propose a method to switch between multi-task learning and single-task learning based on the confidence score of the pseudo-label. The proposed method estimates the confidence level of the pseudo-labels and trains the ASR model by single-task learning using pseudo-labels when the confidence level is high, and by multi-task learning when the confidence level is low.

The rest of this paper is organized as follows. In Section II, we discuss the pseudo-labeling and self-supervised feature leanring and their related works. In Section III, our proposed multi-task learning method is explained. We show our experimental results in Section IV and Section V and conclude this work in Section VI.

II. RELATED WORKS USING UNLABELED SPEECH FOR ASR MODEL TRAINING

A. PSEUDO-LABELING

Pseudo-labeling is a popular approach to utilizing speech data without a human supervision to train an ASR model [30], [31], [32], [33]. In this approach, a seed ASR model is first trained by using labeled speech data. The seed model is used to recognize unlabeled speech data, and the recognition results are used as pseudo-labels for the unlabeled data. Then, an ASR model is re-trained by using the pseudo-labeled speech in addition to the labeled speech. Pseudo-labeling has also been reported to be effective in training end-to-end speech recognition models [19], [20], [21].

The effectiveness of the pseudo-labeling depends on the accuracy of the pseudo-label estimation (i.e., the recognition rate of the seed model). As an extreme example, if the recognition rate is 100%, the pseudo-labeling is equivalent to the supervised learning. However, when applying this approach for dysarthric speech recognition, it should be noted that this method has the risk of training the ASR model using labels containing many errors because estimating labels from dysarthric speech is more difficult than estimating it from normal speech.

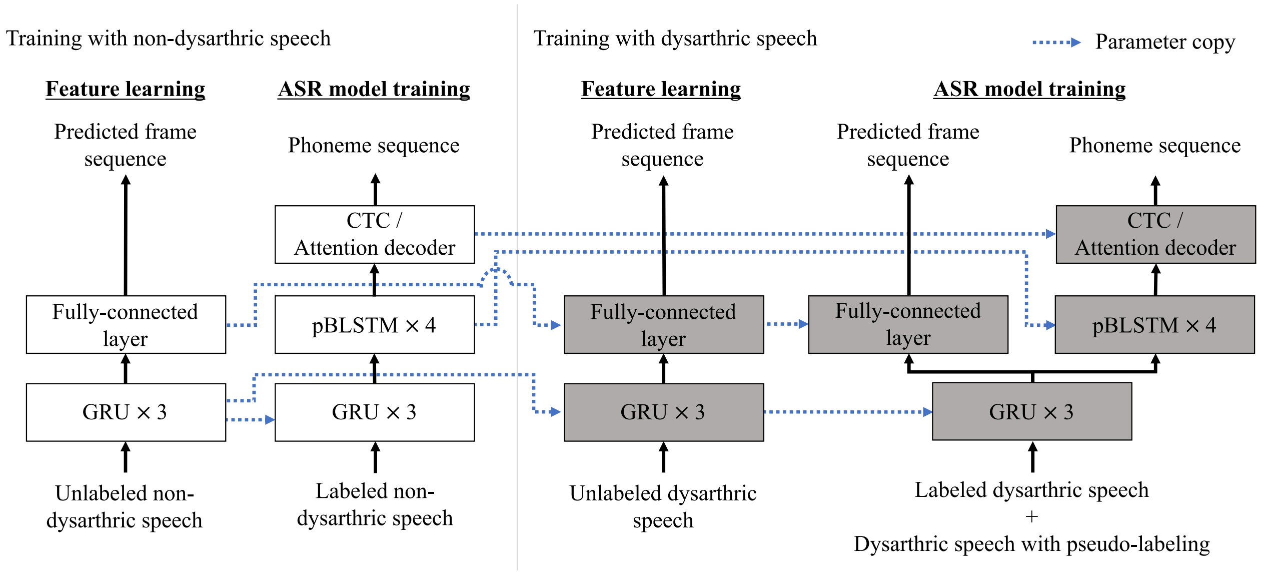


FIGURE 1. Proposed training procedure of the automatic speech recognition (ASR) model.

B. SELF-SUPERVISED FEATURE LEARNING

Self-supervised learning (SSL) is the general term for approaches that use automatically generated labels to train a model without having to manually create labels, and has been widely studied in the fields of natural language processing [34], [35], [36] and computer vision [37], [38], [39], [40]. SSL has also been actively studied in the field of the speech recognition [22], [23], [24], [25], [26], [27], [28], [29], and recent studies [27], [28], [29] reported that the model pre-trained using the SSL with a large amount of unlabeled data can provide promising ASR performance by fine-tuning it with only ten minutes of labeled data.

In the SSL approach, a neural-network model is trained through a pseudo-task that can be generated without human transcription. In this way, the model can learn the representation of the input features, and, therefore, this pre-trained model can be used as a good initial model for the target task. In this paper, we refer to the SSL methods as "*self-supervised feature learning*."

Among the self-supervised feature-learning methods, we employ autoregressive predictive coding (APC) [24]. Although state-of-the-art approaches such as wav2vec 2.0 [27] performed better in the normal speech recognition tasks, our preliminary experiments on dysarthric speech recognition showed that the APC performed better than wav2vec 2.0 (See Section IV-B1).

The APC model consists of a unidirectional recurrent neural network (RNN) and fully-connected layers. The pseudo-task in this method is the prediction of future frames; that is, the APC model is trained so that it predicts the feature vector of the n frame future based on the information of past frames aggregated by the RNN. Specifically, the training is performed so as to minimize the L1-loss between the input sequence $\boldsymbol{x} = (x_1, x_2, \ldots, x_T)$ and the predicted output sequence $\boldsymbol{y} = (y_1, y_2, \ldots, y_T)$, which is expressed as

$$L_{APC} = \sum_{i=1}^{T-n} |x_{i+n} - y_i|. \tag{1}$$

III. PROPOSED METHODS

A. MULTI-TASK LEARNING USING PSEUDO-LABELING AND SELF-SUPERVISED FEATURE LEARNING

In our proposed joint multi-task learning approach, self-supervised feature learning, which is independent of the quality of the pseudo-label, is combined with pseudo-labeling to train an ASR model robustly against the errors contained in the pseudo-labels. Figure 1 shows an overview of the proposed method. First, self-supervised feature learning is conducted using APC and the unlabeled spontaneous speech of a dysarthric speaker. In this process, although spontaneous speech can be collected relatively easily, it should be noted that the amount of available dysarthric speech is still much smaller than normal (non-dysarthric) speech. For this reason, we take a model adaptation approach, where the APC model is pre-trained using a large amount of unlabeled non-dysarthric speech, and the APC model is then fine-tuned using unlabeled dysarthric speech. The APC model consists of a three-layer unidirectional gated recurrent unit (GRU) and a single fully-connected layer.

After conducting the self-supervised feature learning, the ASR model is trained. The ASR model is also pre-trained using a large amount of labeled non-dysarthric speech. Then, the ASR model is fine-tuned using pseudo-labeled speech in addition to the labeled speech of a dysarthric speaker. In [41], it is reported that the ASR performance improved by using pseudo-labeled speech to train an ASR model pre-trained through self-supervised feature-learning (we refer to this approach as "*naive combination*"). In the proposed method, in order to reduce the negative effects of errors in the pseudo-labels, we simultaneously conduct self-supervised feature learning during the training of the ASR model with pseudo-labeled speech. Specifically, the proposed multi-task learning is performed so as to minimize the following equation, which is a weighted sum of the loss function L_{ASR} of the speech recognition task and the loss function L_{APC} of the APC pseudo-task.

$$L = (1 - \lambda)L_{ASR} + \lambda L_{APC} \tag{2}$$

In this study, we use hybrid CTC/attention loss [42], which is proposed for training a combined model of the connectionist temporal classification (CTC) [43] and the encoder-decoder model with an attention mechanism [44], as L_{ASR}. λ is a hyper-parameter that determines the weights of L_{ASR} and L_{APC}.

B. SWITCHING BETWEEN MULTI-TASK AND SINGLE-TASK LEARNING BASED ON THE CONFIDENCE SCORE OF PSEUDO-LABEL

The loss function in Eq. (2) emphasizes the speech recognition loss as the multi-task learning weight λ is closer to 0, while it emphasizes the self-supervised feature learning loss as λ is closer to 1. As mentioned in the previous section, the purpose of the multi-task learning is to reduce the negative effects of the pseudo-label errors. Conversely, if the pseudo-labels have few errors, it is better to train the model with speech recognition loss alone, which is the original objective function. For this reason, we propose to switch between multi-task learning and single-task learning based on the reliability of the pseudo-label. In this method, we define a confidence score as a measure of the degree of errors in pseudo-labels.

For predicting the confidence score of the pseudo-label from a speech utterance, we calculates the mean value of the probability output from the CTC layer as following equation:

$$CS = \frac{1}{T}\sum_{t=1}^{T} \max_i y_i^t, \tag{3}$$

where y_i^t denotes the probability of token i at frame t output from the CTC layer, and T denotes the frame length of the utterance. In this average calculation, frames in which the blank token has the largest probability are not used.

Previous studies have proposed methods of selecting training samples based on the confidence scores of the pseudo-labels, excluding samples with low confidence scores from the training data [45], [46], [47]. In our method, we do not exclude training samples with low confidence scores, but use them with multi-task learning. On the other hand, training samples having true labels (i.e. script-reading data) or pseudo-labels with high confidence scores are used with single-task learning as follows:

$$\lambda = \begin{cases} \lambda_0 & \text{(if Spontaneous speech and } CS < th) \\ 0 & \text{(if Spontaneous speech and } CS \geq th) \\ 0 & \text{(if Script-reading speech),} \end{cases} \tag{4}$$

where λ_0 denotes the weight of the self-supervised feature learning for the multi-task learning, and th denotes the threshold for the confidence score.

IV. EXPERIMENTS ON A JAPANESE DATASET

A. EXPERIMENTAL SETUP

The speech data in our original Japanese dataset was uttered by one Japanese male who has dysarthria due to athetoid cerebral palsy. For labeled script-reading speech, the dysarthric subject read 429 sentences (about one hour) out of 503 phoneme-balanced sentences included in the ATR Japanese speech database [48]. The script-reading speech was divided into 50 sentences for evaluation, 50 sentences for validation, and the rest for training.

For unlabeled spontaneous speech, we recorded a total of 1,460 sentences (about 3 hours) uttered by the subject when he was giving a lecture at a university and when he was reading a newspaper. The spontaneous speech was divided into 76 sentences for evaluation, 59 sentences for validation, and the rest for training. Therefore, the evaluation set for this experiment consisted of 50 sentences of script-reading speech and 76 sentences of spontaneous speech. Although we considered spontaneous speech as unlabeled data in previous sections, the spontaneous speech recorded for our experiments is transcribed, that is, has true labels. The true labels of the spontaneous speech were used for only evaluation and not used for training. When we pre-trained the APC model and the ASR model, we used 660 hours of non-dysarthric speech recorded in the Corpus of Spontaneous Japanese (CSJ) [49].

As input acoustic features, 80-dimensional mel-filterbank features were used. The APC model consisted of a three-layer unidirectional GRU with 512 hidden units for each layer and a single fully-connected layer with 80 output units. The size of the forecasting frame n was set to 1. Adam [50] was used for optimization, the learning rate was 1e-4, and the number of epochs was set to 50.

The output label of the ASR model was defined by phonemes (i.e., a phoneme-level recognition task). The labels consisted of 39 phonemes, unknown symbols, and start and end of sequence symbols (42 dimensions in total). Word error rates or character error rates are commonly used for the evaluation of the normal speech recognition tasks. However, in dysarthric speech recognition tasks, the recognition is often performed at the phoneme level and the phoneme error rate (PER) is used for the evaluation [51], [52]. We conducted phoneme-level speech recognition experiments in order to confirm the performance of purely acoustic models, which is the focus of this study, without the influence of language models (some experimental results on word error rates are given in Appendix).

For the ASR model, the hybrid CTC/attention model [42] was trained using ESPnet toolkit [53]. The shared encoder consisted of a four-layer pyramidal bidirectional long short-term memory (LSTM) with 320 hidden units for each layer. The decoder consisted of a single-layer unidirectional LSTM with 320 hidden units and an attention mechanism, followed by an output layer. The weight of the CTC was set to 0.5 during both training and recognition. Adadelta [54] was used for optimization. The learning rate was set to 1e-8, and the number of epochs was set to 50.

B. RESULTS

1) COMPARISON WITH OTHER MODEL ARCHITECTURES AND WAV2VEC 2.0 AS A SELF-SUPERVISED FEATURE LEARNING METHOD

First, we evaluated the validity of using the APC for self-supervised feature learning by comparing it with wav2vec 2.0. For the baseline systems, we evaluated the LSTM-based model described in Section IV-A and a transformer-based model (2 convolution layers + 12 transformer blocks as encoder and 6 transformer blocks as decoder) which has a similar model architecture to wav2vec 2.0. These baseline models were pre-trained on non-dysarthric speech (660 hours) and then fine-tuned on labeled script-reading dysarthric speech (1 hour). The baselines did not use unlabeled spontaneous speech for training.

For the systems using self-supervised feature learning, we evaluated the APC model consisting of three layers of unidirectional GRU described in Section IV-A, the APC model consisting of three layers of unidirectional LSTM, and wav2vec 2.0 Base model (7 convolution layers + 12 transformer blocks). These systems were pre-trained on non-dysarthric speech (660 hours) and unlabeled spontaneous dysarthric speech (3 hours) and then fine-tuned on labeled script-reading dysarthric speech (1 hour) in the manner described in Section III-A, but neither pseudo-labeling nor multi-task learning was used.

Table 1 shows the PER of each method. As shown in this table, the APC systems, especially the GRU-based APC system, showed the best performance. Wav2vec 2.0 performed better than Baseline (Transformer), which has a similar model architecture, but performed worse than Baseline (LSTM). A previous study [16] reported that state-of-the-art self-supervised learning methods, including wav2vec 2.0, did not outperform a simple LSTM-based model in pathological speech recognition tasks, and our experiments showed similar results. While Baseline (LSTM) and APC systems use relatively small RNN-based models, Baseline (Transformer) and wav2vec 2.0 use large transformer-based models. Therefore, these results may suggest that it is still challenging to train such larger models on small amounts of dysarthric speech, even if they are pre-trained on a large amount of non-dysarthric speech. Similarly, the GRU-based APC system may be better than the LSTM-based APC system because the GRU has fewer parameters than the LSTM. In the following experiments, we use the GRU-based APC for the self-supervised feature learning.

TABLE 1. Comparison between APC model, wav2vec 2.0 and other model architectures.

Method	Self-supervised feature learning	PER [%]
Baseline (LSTM)	no	22.0
Baseline (Transformer)	no	26.4
APC (GRU)	yes	18.7
APC (LSTM)	yes	19.5
Wav2vec 2.0	yes	24.1

TABLE 2. PER [%] of each method without multi-task learning. Self-supervised feature-learning (FL) is the same condition as "APC (GRU)" in Table 1.

Method	PER [%]
Baseline (LSTM)	22.0
Pseudo-labeling (PL)	19.9
Self-supervised feature-learning (FL)	18.7

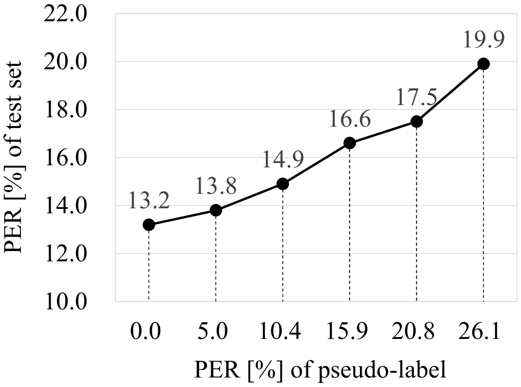


FIGURE 2. The correlation between PER [%] of pseudo-label and PER [%] of test set.

2) EVALUATION WITHOUT MULTI-TASK LEARNING

Table 2 shows the PER of each method without multi-task learning. "Baseline (LSTM)" and "self-supervised feature-learning (FL)" are the same conditions as "Baseline (LSTM)" and "APC (GRU)" in Table 1, respectively.

As shown in Table 2, both pseudo-labeling and feature-learning improved the PER from the baseline, but feature-learning showed better performance than the pseudo-labeling. In this experiment, the PER of the pseudo-label of the training set was 26.1%. In order to analyze the change in speech recognition performance with respect to the accuracy of pseudo-labels, we trained the ASR model by replacing a part of the pseudo-labels with true labels. The experimental results are shown in Figure 2. As shown in this figure, the PER of the test data improved as the pseudo-labels are replaced by the true labels. When all pseudo-labels were replaced with true labels, that is, if the PER for pseudo-labels

TABLE 3. PER [%] of each method combining feature-learning (FL) and pseudo-labeling (PL).

Method	PER [%]
FL + PL (naive combination)	18.1
FL + PL (multi-task learning without switching)	17.4
FL + PL (multi-task learning with switching)	17.1

had been 0%, the PER would improve to 13.2%. These results indicate that because the accuracy of the pseudo-labels of dysarthric speech is worse than that of non-dysarthric speech reported in the previous works, the performance of the pseudo-labeling is limited by the errors in the pseudo-labels.

3) EVALUATION WITH MULTI-TASK LEARNING

Table 3 shows the PER of each method combining feature-learning (FL) and pseudo-labeling (PL). In the naive combination approach [41], the pseudo-label was estimated using the ASR model pre-trained through self-supervised feature-learning (i.e. FL model shown in Table 2),[1] and then, the ASR model was fine-tuned by using script-reading speech and spontaneous speech with the pseudo-label. This method is equivalent to the proposed method when the weight parameter λ in Eq. (2) is set to 0.0. In the method using multi-task learning without switching, the weight parameter λ was set to 0.5. In the method using multi-task learning with switching, the weight parameter λ_0 and the threshold for the confidence score *th* were set to 0.5 and 0.9, respectively.

As shown in Table 3, all three methods showed better performances than methods using PL or FL alone shown in Table 2. When using the multi-task learning, the PERs were improved compared to the naive combination approach. Furthermore, the use of confidence-based switching between multi-task and single-task learning showed slightly lower PER than that without switching.

Comparing Table 3 with Figure 2, we can see that the PER when using multi-task learning without switching (17.4%) is almost the same as the PER of 17.5% when the PER of the pseudo-labels was 20.8% in Figure 2. This result implies that the effect of multi-task learning corresponds to correcting about 20% of the pseudo-label errors (26.1% → 20.8%).

Table 4 shows the effect of the weight parameter for the multi-task learning. Both with and without switching, the lowest PER was obtained when the weight was 0.5, and the PER increased as the weight was decreased or increased from 0.5. Moreover, the use of switching showed lower PERs for all weight parameter settings compared to the PERs without switching. These results indicate that the use of single/multi-task switching had a small but definite effect in this experiment.

Figure 3 shows the relationship between the predicted confidence score and the actual PER of the pseudo-label for each utterance, showing a negative correlation (correlation coefficient was −0.56) between them. Since speech utterances with higher confidence scores tend to have fewer pseudo-label errors, this result indicates that our confidence score can predict the accuracy of the pseudo-label to some extent.

TABLE 4. The effect of the weight parameter for the multi-task learning. Multi-task weight of 0.0 means the naive combination approach without multi-task learning.

Multi-task weight	PER [%]	
	w/o switching	w/switching
0.0 (naive comb.)	18.1	18.1
0.1	17.8	17.5
0.2	17.7	17.4
0.3	17.6	17.2
0.4	17.5	17.2
0.5	**17.4**	**17.1**
0.6	17.4	17.2
0.7	17.6	17.3
0.8	17.7	17.4
0.9	18.4	17.4

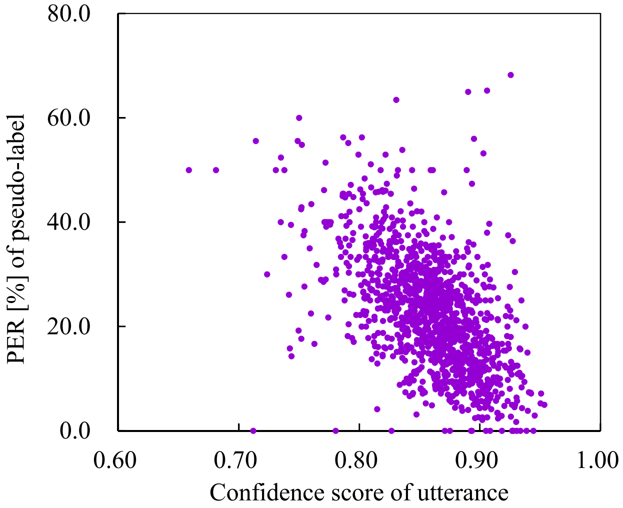


FIGURE 3. The correlation between confidence score and PER of pseudo-label for each utterance. Correlation coefficient was −0.56.

Figure 4 shows the relationship between the predicted confidence score and the actual errors (deletion errors and substitution errors) of the pseudo-label averaged for each phoneme. Similar to the results in Figure 3, a negative correlation (correlation coefficient was −0.69) was observed. However, some long vowels (i.e. "i:" and "e:") had relatively low error rates despite their low confidence scores. This is because when the ASR model recognized those long vowels, it tended to output higher probabilities for the associated short vowels (i.e. "i" and "e"), resulting in relatively low confidence scores for those long vowels. In addition, as shown in this figure, we can see that plosives (i.e. "py" and "by") and fricatives (i.e. "z" and "hy") were difficult to recognize.

V. EXPERIMENTS ON TORGO DATASET

A. EXPERIMENTAL SETUP

We also evaluated our proposed method on an English dataset. As described in the introduction, there are two

[1] The PER of the pseudo-label estimated by FL-based ASR model was 24.3%.

TABLE 5. PERs [%] on TORGO dataset under the overlapping speaker condition. MTL stands for the multi-task learning. "Oracle confidence" refers to the use of the true PER of the pseudo-labels instead of the predicted confidence score in single/multi-task switching. The best result in each speaker is bolded.

Method	Subject								Average
	F01	F03	F04	M01	M02	M03	M04	M05	
Baseline	58.7	52.3	33.5	51.0	49.2	21.7	71.1	72.3	52.0
Pseudo-labeling (PL)	62.9	51.4	31.7	50.6	47.6	21.2	68.2	73.4	51.5
Self-supervised feature-learning (FL)	58.0	43.2	22.5	50.3	49.2	21.9	54.9	60.3	45.3
FL + PL (naive comb.)	56.9	44.7	23.7	48.2	46.1	18.0	54.7	63.2	44.8
FL + PL (MTL w/o switching)	**56.3**	**44.1**	22.9	**47.8**	46.2	17.9	**54.5**	62.4	**44.4**
FL + PL (MTL w/ switching)	56.8	46.0	**22.6**	48.7	**46.0**	**17.8**	55.3	**61.6**	44.7
FL + PL (MTL w/ switching using oracle confidence)	56.3	43.8	22.6	46.4	44.9	18.3	52.7	61.9	43.7

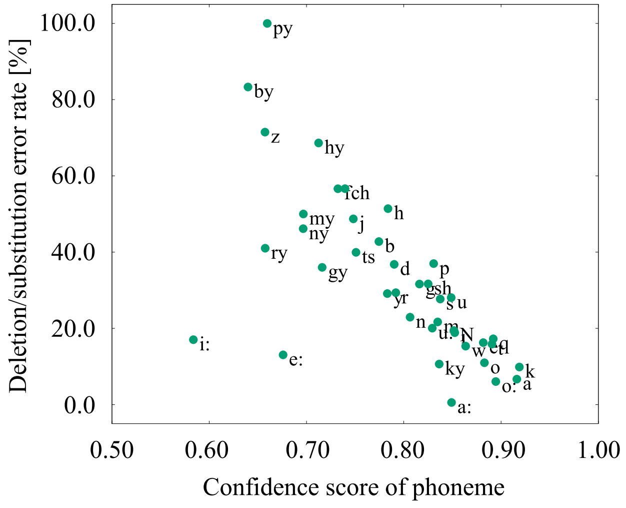


FIGURE 4. The correlation between average confidence score and average deletion/substitution error rate per phoneme.

publicly available representative English dysarthric speech datasets: TORGO dataset [4] and UAspeech dataset [5]; however, we only used TORGO dataset because UAspeech contains only isolated word speech and was difficult to use for self-supervised feature learning.

TORGO dataset contains script-reading speech uttered from eight subjects with either cerebral palsy or amyotrophic lateral sclerosis. Because the TORGO dataest, like the other publicly available datasets, does not have spontaneous speech we focus on in this work, we used 1,803 utterances recorded with array microphones as unlabeled training set for pseudo-labeling and feature learning, and we used 1,958 utterances recorded with a head-mounted microphone as labeled training set. For each subset, the 10% of the training set was used as development set. For the evaluation set, a total of 1,727 utterances, including both speech recorded with array microphones and the head-mounted microphone, were used. In this data setup, the same speaker is included in both the training set and the evaluation set. We call this setup the *"overlapping speaker condition"*. In addition, we conducted experiments on a *"non-overlapping speaker condition"*, where the speaker to be evaluated is not included in the training set. In this condition, each speaker was evaluated individually, and the model was trained by excluding the evaluated speaker's speech from the training data. In both conditions, about 300 hours of non-dysarthric speech from the Librispeech dataset [55] were used for pre-training.

The output labels were defined with 69 phonemes. The other experimental conditions were same as for the Japanese speech recognition task in the previous section. In the method using multi-task learning without switching, the weight parameter λ was set to 0.5. In the method using multi-task learning with switching, the weight parameter λ_0 and the threshold for the confidence score th were set to 0.5 and 0.9, respectively.

B. RESULTS

Table 5 shows the experimental results on the TORGO dataset in the overlapping speaker condition. Focusing on the results using PL and FL alone, both showed PERs averaged over all subjects that were lower than baseline. However, the performance improvement for PL was quite small compared to FL. In this experiment, the PER of the pseudo-label of the training set was 56.5% on average for all subjects. Since the accuracy of the pseudo-labels was worse than that on the experiments using Japanese dysarthric dataset (PER = 26.1%), the pseudo-labeling might not be effective in this experiment.

Three methods combining FL and PL (naive combination, MTL w/o switching, and MTL w/ switching) showed lower average PERs than using FL alone. However, single/multi-task switching did not show significant improvement for the multi-task learning in this experiment. Figure 5 shows the relationship between the predicted confidence score and the actual PER of the pseudo label for each utterance. Compared to the results in Figure 3, the confidence score did not represent the accuracy of the pseudo-labels well (correlation coefficient was −0.43). To see how effectively the switching approach would work if the confidence scores perfectly represented the accuracy of the pseudo-labels, we evaluated the performance by using the true PERs of pseudo-labels as the oracle confidence scores instead of the predicted confidence scores in the single/multi-task switching. The results are shown in the bottom row of Table 5. By using the oracle confidence scores, the switching approach showed the lowest average PER. These results

TABLE 6. PERs [%] on TORGO dataset under the non-overlapping-speaker conditions. The best result in each speaker is bolded.

Method	Subject								Average
	F01	F03	F04	M01	M02	M03	M04	M05	
Baseline	69.7	56.8	42.6	54.1	70.1	25.9	73.5	84.4	59.6
Pseudo-labeling (PL)	64.9	61.7	42.3	63.7	67.6	33.8	73.4	78.8	60.8
Self-supervised feature-learning (FL)	60.6	53.3	33.2	53.1	63.5	23.1	64.1	68.5	52.4
FL + PL (naive comb.)	54.3	52.7	28.7	51.4	57.0	20.5	57.1	68.0	48.7
FL + PL (MTL w/o switching)	**53.3**	51.5	26.0	49.7	55.5	**19.3**	56.2	69.9	47.3
FL + PL (MTL w/ switching)	53.7	**49.6**	**25.7**	**47.9**	**54.8**	19.4	**55.1**	**66.5**	**46.6**

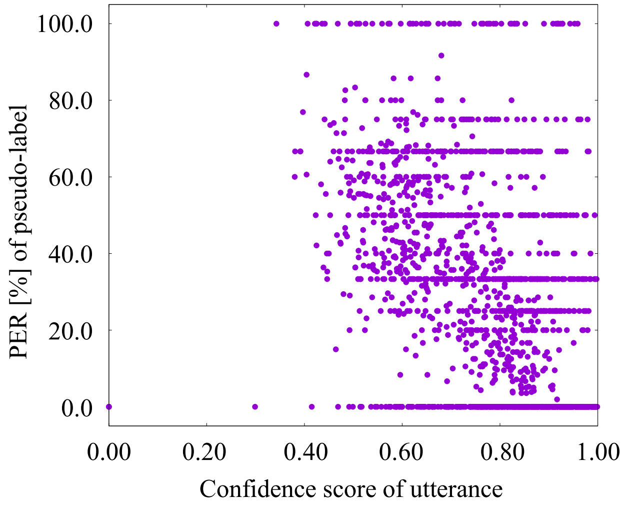


FIGURE 5. The correlation between confidence score and PER of pseudo-label for each utterance in TORGO dataset. Correlation coefficients was −0.43.

TABLE 7. WER [%] of each method without multi-task learning.

Method	WER [%]
Baseline (LSTM)	42.9
Pseudo-labeling (PL)	41.7
Self-supervised feature-learning (FL)	39.4

TABLE 8. WER [%] of each method combining feature-learning (FL) and pseudo-labeling (PL).

Method	WER [%]
FL + PL (naive combination)	37.2
FL + PL (multi-task learning without switching)	36.2
FL + PL (multi-task learning with switching)	36.1

indicate that the effect of the single/multi-task switching was limited because the confidence score did not represent the accuracy of the pseudo-labels well.

Table 6 shows the PERs for the non-overlapping speaker condition. Compared to the overlapping speaker condition shown in Table 5, the overall PERs were higher because the speech of the evaluated speaker was not included in the training set. Nevertheless, similar to the overlapping speaker condition, the FL improved the overall error rate while the effects of the PL were limited. Furthermore, MTL-based approaches between FL and PL showed the best performance for all speakers. These results suggest that the use of the unlabeled speech is also effective even when the speech of the evaluated speaker is not included in the unlabeled speech.

VI. CONCLUSION

In this study, we investigated the use of unlabeled dysarthric speech data with pseudo-labeling and self-supervised feature learning for dysarthric speech recognition and proposed a joint multi-task learning approach with single/multi-task switching based on the confidence score of pseudo-labels.

In our experiments using a Japanese dataset and TORGO dataset, we confirmed that both of pseudo-labeling and self-supervised feature learning were effective to use unlabeled dysarthric speech for training an ASR model. However, because the accuracy of the pseudo-labels of dysarthric speech is worse than that for non-dysarthric speech, the performance of the pseudo-labeling is limited compared to the self-supervised feature learning. In addition, we confirmed that combining the pseudo-labeling and self-supervised feature learning is more effective than using those two methods alone, and furthermore, the multi-task learning worked more effective than the naive combination. However, while the single/multi-task switching worked well in experiments using Japanese dataset, the effect of this approach was limited in experiments using TORGO dataset because the confidence score did not represent the accuracy of the pseudo-labels well. In the future, we will investigate a method to use pseudo-labeling and self-supervised feature learning more effectively.

APPENDIX

EVALUATION ON WORD ERROR RATES

We re-evaluated the experimental results shown in Table 2 and Table 3 using word error rates (WERs). Table 7 and Table 8 show the WERs for the corresponding experimental results. WERs showed higher values than PERs because PER evaluates phoneme errors independently, whereas WER evaluates combinations of phonemes in a word. Nevertheless, similar to the PER results, the best WERs were obtained by the multi-task learning of self-supervised feature learning and pseudo-labeling. For lower WERs, in addition to improving the acoustic model, which is the focus of this study, it would also be necessary to improve the language model.

REFERENCES

[1] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The Microsoft 2017 conversational speech recognition system," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5934–5938.

[2] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, B. Roomi, and P. Hall, "English conversational telephone speech recognition by humans and machines," in *Proc. Interspeech*, Aug. 2017, pp. 132–136.

[3] *Cerebral Palsy: Hope Through Research*, Nat. Inst. Neurological Disorders Stroke, Bethesda, MD, USA, 2009.

[4] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Lang. Resour. Eval.*, vol. 46, no. 4, pp. 523–541, Dec. 2012.

[5] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Proc. Interspeech*, Sep. 2008, pp. 1741–1744.

[6] Y. Jiao, M. Tu, V. Berisha, and J. Liss, "Simulating dysarthric speech for training data augmentation in clinical speech applications," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 6009–6013.

[7] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition," in *Proc. Interspeech*, Sep. 2018, pp. 471–475.

[8] F. Xiong, J. Barker, and H. Christensen, "Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5836–5840.

[9] K. Fujiwara, R. Takashima, C. Sugiyama, N. Tanaka, K. Nohara, K. Nozaki, and T. Takiguchi, "Data augmentation based on frequency warping for recognition of cleft palate speech," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2021, pp. 471–476.

[10] C. Bhat, A. Panda, and H. Strik, "Improved ASR performance for dysarthric speech using two-stage DataAugmentation," in *Proc. Interspeech*, Sep. 2022, pp. 46–50.

[11] J. Shor, D. Emanuel, O. Lang, O. Tuval, M. Brenner, J. Cattiau, F. Vieira, M. McNally, T. Charbonneau, M. Nollstadt, A. Hassidim, and Y. Matias, "Personalizing ASR for dysarthric and accented speech with limited data," in *Proc. Interspeech*, Sep. 2019, pp. 784–788.

[12] Y. Takashima, R. Takashima, T. Takiguchi, and Y. Ariki, "Knowledge transferability between the speech data of persons with dysarthria speaking different languages for dysarthric speech recognition," *IEEE Access*, vol. 7, pp. 164320–164326, 2019.

[13] F. Xiong, J. Barker, Z. Yue, and H. Christensen, "Source domain data selection for improved transfer learning targeting dysarthric speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7424–7428.

[14] R. Takashima, T. Takiguchi, and Y. Ariki, "Two-step acoustic model adaptation for dysarthric speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6104–6108.

[15] A. Hernandez, P. A. Pérez-Toro, E. Noeth, J. R. Orozco-Arroyave, A. Maier, and S. H. Yang, "Cross-lingual self-supervised speech representations for improved dysarthric speech recognition," in *Proc. Interspeech*, Sep. 2022, pp. 51–55.

[16] L. P. Violeta, W. C. Huang, and T. Toda, "Investigating self-supervised pretraining frameworks for pathological speech recognition," in *Proc. Interspeech*, Sep. 2022, pp. 41–45.

[17] M. K. Baskar, T. Herzig, D. Nguyen, M. Diez, T. Polzehl, L. Burget, and J. Černocký, "Speaker adaptation for Wav2vec2 based dysarthric ASR," in *Proc. Interspeech*, Sep. 2022, pp. 3403–3407.

[18] Y. Takashima, T. Takiguchi, and Y. Ariki, "End-to-end dysarthric speech recognition using multiple databases," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6395–6399.

[19] J. Kahn, A. Lee, and A. Hannun, "Self-training for end-to-end speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7084–7088.

[20] Q. Xu, T. Likhomanenko, J. Kahn, A. Hannun, G. Synnaeve, and R. Collobert, "Iterative pseudo-labeling for speech recognition," in *Proc. Interspeech*, Oct. 2020, pp. 1006–1010.

[21] D. S. Park, Y. Zhang, Y. Jia, W. Han, C.-C. Chiu, B. Li, Y. Wu, and Q. V. Le, "Improved noisy student training for automatic speech recognition," in *Proc. Interspeech*, Oct. 2020, pp. 2817–2821.

[22] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[23] D. Jiang, X. Lei, W. Li, N. Luo, Y. Hu, W. Zou, and X. Li, "Improving transformer-based speech recognition using unsupervised pre-training," 2019, *arXiv:1910.09932*.

[24] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," in *Proc. Interspeech*, Sep. 2019, pp. 146–150.

[25] M. Rivière, A. Joulin, P.-E. Mazaré, and E. Dupoux, "Unsupervised pretraining transfers well across languages," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7414–7418.

[26] W. Wang, Q. Tang, and K. Livescu, "Unsupervised pre-training of bidirectional speech encoders via masked reconstruction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6889–6893.

[27] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2020, pp. 12449–12460.

[28] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," 2020, *arXiv:2006.13979*.

[29] W.-N. 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, Language Process.*, vol. 29, pp. 3451–3460, 2021.

[30] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Comput. Speech Lang.*, vol. 16, no. 1, pp. 115–129, Jan. 2002.

[31] H. Y. Chan and P. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2004, p. 737.

[32] F. Wessel and H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 23–31, Jan. 2005.

[33] K. Yu, M. Gales, L. Wang, and P. C. Woodland, "Unsupervised training and directed manual transcription for LVCSR," *Speech Commun.*, vol. 52, nos. 7–8, pp. 652–663, Jul. 2010.

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Jun. 2019, pp. 4171–4186.

[35] T. B. Brown et al., "Language models are few-shot learners," in *Proc. NIPS*, 2020, pp. 1877–1901.

[36] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," in *Proc. ICLR*, Apr. 2020, pp. 1–17.

[37] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, Apr. 2020, pp. 1597–1607.

[38] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. NeurIPS*, Dec. 2020, pp. 9912–9924.

[39] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15745–15753.

[40] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9630–9640.

[41] Q. Xu, A. Baevski, T. Likhomanenko, P. Tomasello, A. Conneau, R. Collobert, G. Synnaeve, and M. Auli, "Self-training and pre-training are complementary for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 3030–3034.

[42] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1240–1253, Dec. 2017.

[43] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.

[44] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 4945–4949.

[45] S. Li, Z. Wei, J. Zhang, and L. Xiao, ‘‘Pseudo-label selection for deep semi-supervised learning,’’ in *Proc. IEEE Int. Conf. Prog. Informat. Comput. (PIC)*, Dec. 2020, pp. 1–5.

[46] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, ‘‘In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning,’’ in *Proc. ICLR*, May 2021, pp. 1–20.

[47] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki, ‘‘FlexMatch: Boosting semi-supervised learning with curriculum pseudo labeling,’’ in *Proc. NeurIPS*, Dec. 2021, pp. 1–12.

[48] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, ‘‘ATR Japanese speech database as a tool of speech recognition and synthesis,’’ *Speech Commun.*, vol. 9, no. 4, pp. 357–363, Aug. 1990.

[49] K. Maekawa, ‘‘Corpus of spontaneous Japanese: Its design and evaluation,’’ in *Proc. ISCA IEEE Workshop Spontaneous Speech Process. Recognit.*, Apr. 2003, pp. 7–12.

[50] D. P. Kingma and J. L. Ba, ‘‘Adam: A method for stochastic optimization,’’ in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[51] Y. Lin, L. Wang, S. Li, J. Dang, and C. Ding, ‘‘Staged knowledge distillation for end-to-end dysarthric speech recognition and speech attribute transcription,’’ in *Proc. Interspeech*, Oct. 2020, pp. 4791–4795.

[52] Z. Qian and K. Xiao, ‘‘A survey of automatic speech recognition for dysarthric speech,’’ *Electronics*, vol. 12, no. 20, p. 4278, Oct. 2023.

[53] 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*, Sep. 2018, pp. 2207–2211.

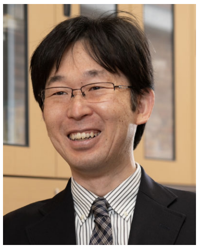
[54] M. D. Zeiler, ‘‘ADADELTA: An adaptive learning rate method,’’ 2012, *arXiv:1212.5701*.

[55] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, ‘‘Librispeech: An ASR corpus based on public domain audio books,’’ in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.

RYOICHI TAKASHIMA (Member, IEEE) received the B.E., M.E., and Dr.-Eng. degrees in computer science from Kobe University, in 2008, 2010, and 2013, respectively. From 2013 to 2018, he was a Researcher with Hitachi Ltd., Tokyo, Japan, and from 2016 to 2018, he was on loan to the National Institute of Information and Communication Technology (NICT), Kyoto, Japan. He is currently an Associate Professor with Kobe University. His research interests include machine learning and signal processing. He is a member of IEICE, IPSJ, and ASJ.

YUYA SAWA received the B.E. and M.E. degrees in system informatics from Kobe University, in 2020 and 2022, respectively. His research interests include speech recognition and assistive technologies for people with articulation disorders.

RYO AIHARA (Member, IEEE) received the B.Eng., M.Eng., and Ph.D. degrees from Kobe University, Japan, in 2012, 2014, and 2017, respectively. He is currently the Head Researcher of the Information Technology Research and Development Center, Mitsubishi Electric Corporation, Kamakura, Japan. His research interests include signal processing and machine learning applied to speech, audio, and image. He is a member of the Acoustic Society of Japan. He was a recipient of Japan Society for the Promotion of Science Research Fellowship for Young Scientists (DC1) from 2014 to 2017.



TETSUYA TAKIGUCHI (Member, IEEE) received the M.Eng. and Dr.-Eng. degrees in information science. He was a Researcher with Tokyo Research Laboratory, IBM Research. From 2004 to 2016, he was an Associate Professor with Kobe University, where he has been a Professor, since 2016. From May 2008 to September 2008, he was a Visiting Scholar with the Department of Electrical Engineering, University of Washington. From March 2010 to September 2010, he was a Visiting Scholar with the Institute for Learning and Brain Sciences, University of Washington. From April 2013 to October 2013, he was a Visiting Scholar with Laboratoire d'InfoRmatique en Image et Systèmes d'information, INSA Lyon. His research interests include speech, image, and brain processing, and multimodal assistive technologies for people with articulation disorders. He is currently a member of IEICE, IPSJ, and ASJ.

YOSHIE IMAI received the B.Eng., M.Eng., and Ph.D. degrees from Chiba University, Japan. She is currently a Senior Researcher and a Group Manager with the Information Technology Research and Development Center, Mitsubishi Electric Corporation, Kamakura, Japan. Before joining Mitsubishi Electric, she was with the Research and Development Center, Toshiba Corporation, Kawasaki, Japan. Her research interests include computer vision, image processing, image quality evaluation, and machine learning.