Decoding Imagined Speech Based on Deep Metric Learning for Intuitive BCI Communication

Dong-Yeon Lee, Minji Lee, and Seong-Whan Lee, *Fellow, IEEE*

***Abstract*— Imagined speech is a highly promising paradigm due to its intuitive application and multiclass scalability in the field of brain-computer interfaces. However, optimal feature extraction and classifiers have not yet been established. Furthermore, retraining still requires a large number of trials when new classes are added. The aim of this study is (i) to increase the classification performance for imagined speech and (ii) to apply a new class using a pretrained classifier with a small number of trials. We propose a novel framework based on deep metric learning that learns the distance by comparing the similarity between samples. We also applied the instantaneous frequency and spectral entropy used for speech signals to electroencephalography signals during imagined speech. The method was evaluated on two public datasets (6-class Coretto DB and 5-class BCI Competition DB). We achieved a 6-class accuracy of 45.00 ± 3.13% and a 5-class accuracy of 48.10 ± 3.68% using the proposed method, which significantly outperformed state-of-the-art methods. Additionally, we verified that the new class could be detected through incremental learning with a small number of trials. As a result, the average accuracy is 44.50 ± 0.26% for Coretto DB and 47.12 ± 0.27% for BCI Competition DB, which shows similar accuracy to baseline accuracy without incremental learning. Our results have shown that the accuracy can be greatly improved even with a small number of trials by selecting appropriate features from imagined speech. The proposed framework could be directly used to help construct an extensible intuitive communication system based on brain-computer interfaces.**

***Index Terms*— Imagined speech, instantaneous frequency, spectral entropy, deep metric learning, brain-computer interface.**

Manuscript received January 13, 2021; revised April 24, 2021 and June 15, 2021; accepted July 9, 2021. Date of publication July 13, 2021; date of current version July 20, 2021. This work was supported in part by the Institute for Information and Communications Technology Promotion (IITP) Grant through the Korean Government (Development of BCI-based Brain and Cognitive Computing Technology for Recognizing User's Intentions using Deep Learning) under Grant 2017-0-00451 and in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant through the Korean Government (MSIT) (Artificial Intelligence Graduate School Program (Korea University)) under Grant 2019-0-00079. *(Dong-Yeon Lee and Minji Lee contributed equally to this work.) (Corresponding author: Seong-Whan Lee.)*

Dong-Yeon Lee and Minji Lee are with the Department of Brain and Cognitive Engineering, Korea University, Seongbuk-gu, Seoul 02841, Republic of Korea (e-mail: dongyeon_lee@korea.ac.kr; minjilee@korea.ac.kr).

Seong-Whan Lee is with the Department of Artificial Intelligence, Korea University, Seongbuk-gu, Seoul 02841, Republic of Korea (e-mail: sw.lee@korea.ac.kr).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNSRE.2021.3096874.

Digital Object Identifier 10.1109/TNSRE.2021.3096874

I. Introduction

BRAIN-COMPUTER interfaces (BCIs) refer to controlling the external devices by detecting brain signals [1]. Previous studies have mainly used external stimuli such as visual stimuli, but recent research has focused on extensible and intuitive paradigms for practical application [2]. The BCI paradigm is divided into an exogenous paradigm that measures brain waves responding to external stimuli and an endogenous paradigm that measures spontaneous electroencephalography (EEG) signals according to user intention [3]. The exogenous paradigm has the advantage of high classification accuracy but has the disadvantage of requiring external devices [4]. On the other hand, the endogenous paradigm does not require external stimuli but does not perform as well as the exogenous paradigm [2]. Because external devices are not required, many endogenous paradigms have recently been studied in the BCI field. Traditionally, motor imagery (MI) was the most studied endogenous paradigm [5], [6]. MI refers to measuring the brain waves generated when we imagine the intention of movement [7]. Therefore, the number of classes is limited because of the imagination of body movement (i.e., right hand, left hand, and foot). Therefore, it is not suitable and intuitive for communication systems that require a large number of available classes [7], [8].

Imagined speech has recently been studied as an intuitive paradigm [2]. Brain waves generated by imagining pronunciation without movement of the articulators are measured [9]. This paradigm is particularly suitable for building communication systems due to its intuitiveness. However, the performance is still not as high as that of other paradigms. Nevertheless, imagined speech has multiclass scalability [2], thus showing the possibility of building an extensible BCI system. When increasing the number of classes, it is inefficient to repeat the whole training process. The process of acquiring and retraining data is also a time-consuming task. García-Salinas *et al.* [10] proposed a framework for an extensible BCI system. They used a bag of features (BoF) feature extractor and a naive Bayes classifier to classifying 5 imagined words. Then, after constructing a BoF with 4 words, transfer learning was performed on the new word to explore whether it could be composed of the previous features. This study showed the possibility of predicting a new word through a pretrained BoF feature extractor. However, this study requires many trials for new words to train the classifier, and

there is performance degradation compared to the baseline accuracy.

The brain signals during the imagined speech were associated with speech signals. In a paper by Coffey *et al.* [11], when receiving audio stimuli, the magnetoencephalography (MEG) signals appeared analogous to the envelope of the speech signals, and there was a slight delay. In a paper by Watanabe *et al.* [12], a similar trend in EEG signals appeared when receiving audio stimuli and when performing imagined speech. Similarly, there was a little delay when performing an imagined speech. Thus, it is considered that EEG oscillations during imagined and perceived speech are synchronized with the envelope of speech signals.

In this study, we proposed a novel framework to increase the classification performance during imagined speech and the number of classes even with a small number of EEG trials for an intuitive, extensible BCI system. We used deep metric learning, which refers to a method of distance training by comparing the similarity between samples using deep learning [13]. Siamese neural networks are a kind of deep metric learning network and have the advantage of reducing the dimensions of high-dimensional signals [14], [15]. The instantaneous frequency and spectral entropy were used for extracting suitable EEG features during imagined speech. The instantaneous frequency can extract the frequency component of signals that changes over time [16], and the spectral entropy reveals the waveform of the signal in white noise [17]. These features have been widely used in the field of speech recognition [18], [19].

Additionally, the incremental learning method was used to check multiclass scalability in the proposed framework. This method uses a pretrained classifier to classify new data that have not been encountered [20], [21]. In other words, this method aims to classify new data while not losing its original capabilities [22]. In this study, we hypothesized that the newly added data were a new class. At this time, the class used for pretraining is called the base class, and the new class is called the novel class. Specifically, a small number of EEG trials were used whenever the number of classes increased using incremental learning. To the best of our knowledge, there is no approach to classify EEG signals during imagined speech using conditional spectral moments of the time-frequency distribution based on a deep metric learning approach. The proposed framework improves the performance of imagined speech systems and classifies the novel class while maintaining the original capabilities. These results show the possibilities for an intuitive BCI system using imagined speech.

II. Related Works

Imagined speech is the BCI paradigm that is drawing attention for intuitive and multiclass scalability [2]. Many studies are reported in two directions; (i) extracting suitable features for imagined speech and (ii) building optimal classifiers.

A. Feature Extraction

The common spatial patterns (CSP) is the most widely used method for feature extraction in MI paradigm. This has already been demonstrated as a distinguishing feature, such as left hand, right hand, during MI [23]. Imagined speech is a recently emerging endogenous paradigm and it has not yet been known which feature extraction method is suitable. In this respect, CSP is commonly used in imagined speech because it is an endogenous paradigm such as MI [7]. This method finds the optimal spatial filters, which maximize the variance of the EEG signals of one class and minimize those of the other class, using covariance matrices [24]. Dasalla *et al.* [23] used CSP and a support vector machine (SVM) to classify binary imagined vowels. The minimum classification accuracy was 68%, and the maximum was 78%. Lee *et al.* [7] attempted to classify 13 classes (12 imagined words with the resting state) to highlight the advantage of imagined speech that many classes can be used compared to the existing paradigm such as MI. A random forest (RF) model showed an average classification accuracy of 20.4% for 13 classes, including the resting state. The CSP algorithm has the advantage of reducing the dimension of data and increasing the distance between different classes but has the disadvantage of being optimized in binary classes [25]. Therefore, many studies have been performed to extract new features that are suitable for multiclass classification and can take advantage of the features of speech itself.

Coretto *et al.* [26] used the discrete wavelet transform (DWT) as a feature extractor and an RF as a classifier to classify 5 vowels and 6 words during imagined speech. As a result, the 5-class vowel accuracy was 22.72% and the 6-class word accuracy was 19.60%. García-Salinas *et al.* [10] reported that 5-class classification accuracy had an average accuracy of 68.9% using the BoF approach. However, they did not give a random cue for imagined speech. This is a limitation in that stimulation should be presented randomly to exclude habituation effects in the field of BCI [27]. They also admit that the high performance is due to this characteristic of the paradigm. Nevertheless, it is implied that the classification performance of the imagined speech can be improved by selecting suitable features [28].

In Dash *et al.* [29], they proposed the forward selection algorithm using spatial selectivity of MEG signals. The aim of this study is to minimize the number of sensors for neural speech decoding. In specific, the optimal sensor subset of the whole feature set was selected by removing redundant or irrelevant MEG features from the data during speech decoding. As a result, the authors obtained higher accuracy using only nine sensors located near the Broca's area compared to using all channels.

B. Classifier Training

Since suitable features for imagined speech have not yet been found, many studies are being conducted to improve performance by automatically extracting features based on deep learning. Saha *et al.* [30] extracted features using a convolutional neural network (CNN) and a long-short term memory (LSTM) network in parallel and then concatenated these features. The channel cross-covariance was used as the

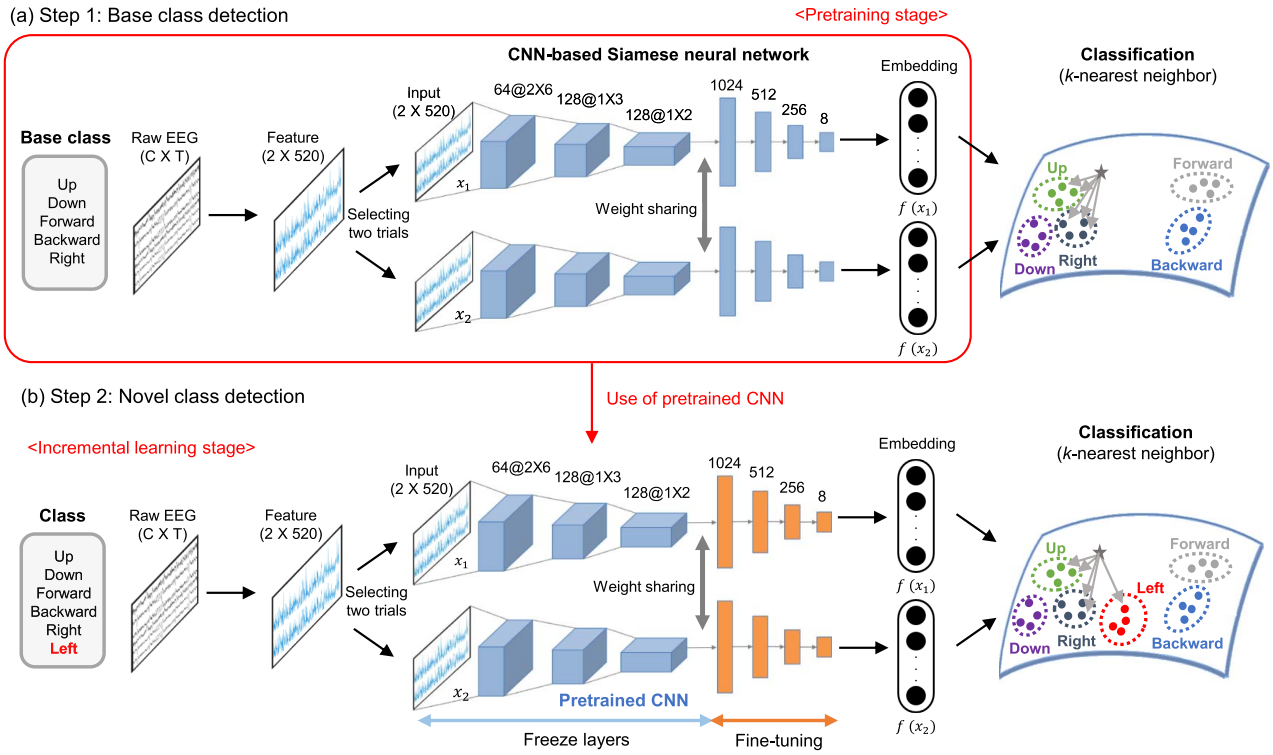

Fig. 1. Proposed framework for classifying EEG signals during imagined speech. (a) In step 1, the classifier was trained using the base class ('up,' 'down,' 'forward,' 'backward,' and 'right') without incremental learning. Each raw EEG data point consists of C channels and T time points. The features indicate the instantaneous frequency and spectral entropy extracted from the raw EEG signals. x_1 and x_2 denote two trials chosen at random regardless of class. $f(x_1)$ and $f(x_2)$ denote the reduced embeddings extracted through the Siamese neural network. Finally, dimensionally reduced embeddings are classified through the k-NN classifier. One of the test trials is marked with a star. (b) In step 2, a novel class ('left') was detected using incremental learning based on the pretrained classifier. The pretraining stage is the same as that in step 1. The weight in this layer (called freeze layers) does not change when the network pretrained using five base classes is retrained to classify the novel class. But, the fully connected layers of the Siamese neural network are fine-tuned.

network input, and the network was trained using a deep autoencoder. In summary, they tried to use this network to extract spatiotemporal information. This method achieved averaged accuracy of 77.9% for 2-class classification. In their recent work [31], they refined the previous method slightly, replacing the LSTM networks with temporal CNNs to evaluate performance. This method achieved an average 2-class classification accuracy of 83.42%, which was 5.52% higher than that of the previous method.

Furthermore, Dash *et al.* [32] decoded five imagined phrases using MEG signals. The authors used CNN applied on the spatial, spectral, and temporal features in terms of scalograms of the neuromagnetic signals. As a result, they achieved an average accuracy of up to 93% during imagined speech. However, although MEG signals have a high-temporal resolution, this is limited to practical use in real BCIs due to their high price, large size and weight [29].

Cooney *et al.* [33] used independent component analysis with Hessian approximation preconditioning to eliminate electrooculography signals. Using a CNN, it achieved a 32.35% accuracy in 5-class imagined vowel classifications. In their recent study [34], they used the same preprocessing method and then classified word-pairs by applying DeepConvNet [35] and ShallowConvNet [35], which are commonly used in EEG signal classification. These methods achieved a 62.37% classification accuracy for word pairs (2-class classification). However, these studies were not aimed at finding suitable features for imagined speech. They simply designed the optimal structure of the network for high performance so that deep learning could learn its features. Furthermore, they still have the limitation of training classifiers again when applying the new data for multiclass scalability.

III. Methods

A. Overall Framework

The proposed framework consists of two steps. The first step is to classify the base class, and the second step is to detect the novel class. The overall flow of the two steps is the same, but in the second step, fine-tuning is performed using the weights of the network pretrained in the first step. Each step consists of feature extraction and classification. First, the instantaneous frequency and spectral entropy are extracted from the raw signals and then merged into 2-dimensional data. Then, the extracted data are input into the Siamese neural network to learn the distance between the training sets. After training the Siamese neural network, the initial inputs are reduced to an 8-dimensional embedding. These embeddings train the k-nearest neighbors (k-NN) classifier and can classify the class during imagined speech.

Fig. 1 shows a detailed description of the proposed framework. In the Coretto DB [26], there are six words such as

'up,' 'down,' 'forward,' 'backward,' 'right,' and 'left'. Here, we suppose 'left' as the novel class. In step 1, the classifier was learned using the base class. This is the baseline classification without incremental learning. We used 5-fold cross-validation, so 80% of the data is used for training and 20% of the data used for testing. The raw EEG signals converted to the instantaneous frequency and spectral entropy. These signals are used as input into the Siamese neural network to reduce the dimensions. Finally, dimensionally reduced embeddings are classified using the k-NN classifier. When the test trial is placed in an embedding space through the trained Siamese neural network, the class is determined by considering the five closest instances in this embedding space. In step 2, a novel class was detected using the pretrained classifier shown in step 1. In incremental learning stage, 20% of the base class and 20% of the novel class 'left' are trained using pretrained classifier. Data from 20% of the base class is selected from the data used in step 1. The weight in freeze layers does not change when the network pretrained using five base classes is retrained to detect the novel class. But, the fully connected layers of the Siamese neural network are fine-tuned. From the perspective of overall training, the base class used 80% of the data for training, but only 20% of the data in the novel class are used.

B. Feature Extraction

1) Instantaneous Frequency: The instantaneous frequency of signals represents the average of the frequencies present in the signal as it evolves [36]. It is estimated as the first conditional spectral moment of the time-frequency distribution of the input signals. It is calculated as follows:

$$f_{inst}(t) = \frac{\int_0^\infty f P(t, f) df}{\int_0^\infty P(t, f) df} \tag{1}$$

where $P(t, f)$ denotes the power spectrum of time t and frequency f. Since the instantaneous frequency is calculated through one channel, the EEG signals of each channel are concatenated into one dimension.

2) Spectral Entropy: The spectral entropy of a signal is a measure of its spectral power distribution. The entropy represents the uniformity of the spectral power distribution. Hence, when signals other than white noise occur, it generates a small entropy. On the other hand, when only white noise is present, it produces the greatest entropy [37]. The concept is based on the Shannon entropy (information entropy) in information theory [18]. The spectral entropy equations arise from the equations for the power spectrum and probability distribution for a signal. For a signal $x(n)$; $(n = 1, 2, \ldots, N)$, the power spectrum is $s(m) = |X(m)|^2$; $(m = 1, 2, \ldots, N)$, where $X(m)$ is the discrete Fourier transform of $x(n)$ and N is the total number of frequency points. The probability distribution $p(m)$ is calculated as follows:

$$p(m) = \frac{s(m)}{\Sigma_i^N s(i)} \tag{2}$$

The spectral entropy H is as follows:

$$H = -\Sigma_{m=1}^N p(m) log_2 p(m) \tag{3}$$

TABLE I
NEURAL ARCHITECTURE FOR DEEP METRIC LEARNING

Layer	Input	Output	Kernel
Conv1	$2 \times 520 \times 1$	$1 \times 260 \times 64$	2×6
Maxpool1	$1 \times 260 \times 64$	$1 \times 130 \times 64$	1×2
Conv2	$1 \times 130 \times 64$	$1 \times 65 \times 128$	1×3
Maxpool2	$1 \times 65 \times 128$	$1 \times 33 \times 128$	1×3
Conv3	$1 \times 33 \times 128$	$1 \times 17 \times 128$	1×2
Maxpool3	$1 \times 17 \times 128$	$1 \times 9 \times 128$	1×2
Flatten	$1 \times 9 \times 128$	1152	-
Fc1	1152	1024	-
Fc2	1024	512	-
Fc3	512	256	-
Fc4	256	8	-

The normalized equation is as follows:

$$H_n = -\frac{\Sigma_{m=1}^N p(m) \log_2 p(m)}{\log_2 N} \tag{4}$$

where $\log_2 N$ is the maximum spectral entropy of white noise, which is evenly distributed over the frequency domain. If the time-frequency power spectrogram $S(t, f)$ can be found, the probability distribution is as follows:

$$p(m) = \frac{\Sigma_t S(t, m)}{\Sigma_f \Sigma_t S(t, f)} \tag{5}$$

When calculating the instantaneous spectral entropy, the probability distribution at time t is as follows:

$$p(t, m) = \frac{S(t, m)}{\Sigma_f S(t, f)} \tag{6}$$

Instantaneous spectral entropy at time t is as follows:

$$H(t) = -\Sigma_{m=1}^N p(t, m) log_2 p(t, m) \tag{7}$$

Similar to the instantaneous frequency, since the spectral entropy is calculated using one channel value, signals in each channel are concatenated in one dimension. Afterward, the instantaneous frequency and spectral entropy, each composed of 1 dimension, are reconstructed into a signal with 2 dimensions.

3) Feature Extraction Using Deep Metric Learning: It takes much time to acquire a large number of EEG trials. In addition, EEG features extracted through the instantaneous frequency and spectral entropy still have a large dimension. To solve these problems, we used Siamese neural networks for deep metric learning. Siamese neural networks can reduce overfitting by reducing the dimension of the signals and learning the distance between the reduced embeddings. The neural architecture used for training is shown in Table I. This architecture was used as an encoder to reduce the dimension of the data and was trained using Siamese neural networks. In specific, this type of network learns two inputs by randomly selecting them regardless of their class in the training set [14], [15]. Each input obtains a reduced embedding through each network and this network has the same structure and parameters. Randomly selected inputs are denoted by x_1 and x_2. The extracted embeddings through the network are also denoted by $F(x_1)$ and $F(x_2)$. Here, each parameter describes $x_1 \in R^{2\times520}$, $x_2 \in R^{2\times520}$, $F(x_1) \in R^8$, and $F(x_2) \in R^8$. 2×520 comes from Conv1, and 8 comes from Fc4 shown in Table I.

The two extracted embeddings obtained through the network are trained to be located close to each other if the classes are the same and far apart from each other if the classes are different [25]. A Siamese neural network requires two inputs and a corresponding label at each iteration. The information structure is $[x_1, x_2, y]$. y is a label indicating whether the two inputs are of the same class ($y = 1$) or different classes ($y = 0$). The two extracted embeddings are learned through the Euclidean distance and the distance between the two embeddings is calculated as follows:

$$D(x_1, x_2) = \| F(x_1) - F(x_2) \|_2 \tag{8}$$

Siamese neural network learns the contrastive loss based on the above distance, and the loss function is as follows:

$$L = \frac{1}{2} y D^2 + \frac{1}{2}(1 - y) max(m - D, 0)^2 \tag{9}$$

where m denotes a margin ($m > 0$). The purpose of this parameter is to move two inputs farther than the set value when they belong to different classes [13]. A Siamese neural network was originally proposed to use deep learning while avoiding overfitting when the number of training sets is very small as with one-shot learning [14]. By using two inputs, the number of training sets can be increased by the number of combinations of inputs. In this study, we used as many combinations of inputs as much as possible to increase performance. We also tried to obtain a reduced embedding rather than predicting whether two inputs are the same or different through the Siamese neural network.

C. Classification Using the k-NN Algorithm

Feature extracted embeddings were obtained from the high dimensional original data through the instantaneous frequency, spectral entropy, and Siamese neural networks. We used the k-NN algorithm, which works according to the proximity principle of the instance, to classify embeddings into classes [15]. This determines a class by referring to the class of the k-nearest instances in the arbitrary embedding space. In other words, if instances are labeled, the labels for unclassified instances can be decided by observing the labels of their nearest neighbors [38]. Since a Siamese neural network also learns the distance between data in arbitrary embedding space, the k-NN algorithm, which classifies based on distance, is more suitable for classification than other methods. In other words, it is appropriate to use k-NN algorithm as the distance-based classification method because embeddings extracted through the Siamese neural network have distance information. While k-NN has the advantage of being simple and robust, it also has the disadvantage of having a lot of computational burdens caused for training the model [39]. Moreover, the accuracy of the high-dimensional data is degraded due to the curse of dimensionality [40]. In this paper, we tried to compensate for the drawbacks as much as possible by reducing the dimension of the data through a Siamese neural network.

1) Base Class Detection: The base class refers to an existing class to which no new words have been added, and the framework learned from the base class is set as the pretrained framework. The base class detection framework is shown in Fig. 1(a). First, the instantaneous frequency and spectral entropy of raw EEG signals are input into the Siamese neural network. Two inputs are required to be used in the Siamese neural network. Therefore, two random samples are selected from the complete training set regardless of the class. Each sample trains a CNN branch and both branches have the same structure and parameters. If the classes of the two datasets are the same, the embedding extracted through the Siamese neural network is learned to be close to each other, and in case the classes of the two datasets are different, the embedding is learned to be distant. As a result, the Siamese neural network learns the distances between samples in the specific space. By using Siamese neural networks over high-dimensional data, we can obtain a dimensionally reduced embedding that contains important information related to the distance. For classification, the k-NN algorithm was used; it predicts the classes of test data through the closest k instances in a specific space.

2) Novel Class Detection: Novel class detection means classifying newly added classes in addition to the previously learned classes. We conducted novel class detection to show that the proposed method is robust to the extensibility of imagined speech. Since an imagined speech BCI system should be able to incrementally learn about the new class while avoiding catastrophic forgetting [41]. Traditional neural network approaches need to retrain the whole network and add a large amount of data when increasing the number of classes. In addition, most of the incremental learning methods in the BCI system cannot preserve the original capabilities, which refers to the ability to perform existing tasks excluding new tasks when performing incremental learning [42]. To improve these problems, we propose a framework that can maintain the original capabilities while using a small amount of data.

The overall flowchart for incremental learning is shown in Fig. 1(b). First, the Siamese neural network is pretrained using the base class. Next, when a new class is added, fine-tuning is performed with some of the data used in the base class and the data of the novel class. The fine-tuning method is the basic method used when performing incremental learning. The network structure is the same as the structure used for pretraining, and the pretrained weight is used as the initial weight. Based on this weight, only the weights of the fully-connected layers are adjusted. We hypothesized that the convolution layer can extract optimal EEG characteristics through pretraining. Thus, we fine-tune only the fully connected layers. It is also assumed that fine-tuning only the fully connected layers prevents overfitting and destroying the trained layer.

D. Training and Evaluation Scheme

We evaluated the performance on each subject using 5-fold cross-validation. We also used the same number of training sets per class. The Siamese neural network was trained using a contrastive loss function and the Adam optimizer. We set the gradient decay factor to 0.9 and the squared gradient decay factor to 0.99. The rectified linear units (ReLU) function was used after each convolution layer and fully-connected layer.

We also set the margin $m = 0.5$ and the learning rate to 1×10^{-4}. During pretraining, we set the batch size to 500 for each iteration, and train 1000 iterations. In the incremental learning phase, we set the batch size to 100 for each iteration, and train 500 iterations. In the k-NN classifier, the number of nearest neighbors is set to 5 ($k = 5$).

E. Statistical Analysis

We performed a statistical analysis to determine whether the difference between the accuracy of the proposed method and different methods is significant. One-way analysis of variance (ANOVA) was used, and post-hoc analysis was performed using a paired t-test. All the significance levels were set to $\alpha = 0.05$ with Bonferroni correction.

IV. Experimental Results

A. Data Description

We evaluated the classification performance on two public datasets for imagined speech. The detailed comparison is shown in Table II. The first is the open dataset from Coretto DB [26]. It consists of imagined speech and overt speech in EEG signals for a total of 15 subjects. We only used EEG signals during imagined speech. Six electrodes (F3, F4, C3, C4, P3, and P4) were used, and the reference and ground were placed over the left and right mastoids, respectively. In addition, some of the electrodes are located in Wernicke's area, which plays an important role in language processing [28], [43]. The classes consist of 5 Spanish vowels and 6 Spanish words. Words are more intuitive to control external devices than vowels. To evaluate the extensibility of the words, our study only used signals for 6 words: 'arriba' (up), 'abajo' (down), 'derecha' (right), 'izquierda' (left), 'adelante' (forward), and 'atrás' (backward). These words were chosen to intuitively control external devices.

The second dataset is the BCI Competition DB from Track 3 in the 2020 International BCI Competition (https://osf.io/pq7vb/) [44]. This dataset measured EEG signals during imagined speech and consisted of 15 subjects. In total, 64 electrodes were used, and the ground and reference electrodes were placed on Fpz and Fcz, respectively. The five words consist of 'hello', 'help me', 'stop', 'thank you', and 'yes', which are useful commands for patients.

The two datasets consist of a total of 6 classes and 5 classes, and each class includes 40 and 70 trials, respectively. In base class detection without incremental learning, since we performed 5-fold cross-validation, we used 32 and 8 trials in each class in training and testing from Coretto DB [26], respectively. Similarly, in BCI Competition DB [44], 56 and 14 trials were used for training and testing, respectively. When applying to incremental learning, the training and test data were divided for a rehearsal method. This method is used to adapt some of the data that were used during pretraining to maintain the original capabilities when performing incremental learning [20], [41], [45]. In the rest of the classes (base class) except one class (novel class), pretraining is performed using 32 and 56 trials for each class. The novel class is learned by using 8 and 14 trials for each class of the pretraining set and 8 and 14 trials of the novel class together through the incremental learning stage (Supplementary Fig. S1). In the Siamese neural network, data of the same class or different classes are used at the same time as inputs, so a small amount of data of classes other than the novel class is stored and used. Then, unseen 8 and 14 trials were tested in both datasets. In other words, 32 and 56 trials were used to the pretraining stage in the base class detection without incremental learning, whereas only 8 and 14 trials were used for training when applying incremental learning as the novel class.

TABLE II
Comparison Between Two Public Datasets

Dataset	Coretto [26]	BCI Competition [44]
# of EEG channels	6	64
Sampling rate	128 Hz	256 Hz
# of subjects	15	15
Stimuli presentation	Visual and audio stimuli	Visual and audio stimuli
Class	Arriba (up), Abajo (down), Derecha (right), Izquierda (left), Adelante (forward), Atrás (backward)	Hello, Help me, Stop, Thank you, Yes
# of trials	40	70
Cue length	4 sec	2 sec

B. Baseline Classification of Base Class

We trained all the classes to evaluate whether the proposed framework can improve the performance of imagined speech. Fig. 2 illustrates the confusion matrix of the two datasets. The obtained accuracies of all the words were higher than the chance level. In the Coretto DB [26], the 6-class classification accuracy using our proposed pretraining framework shows an average accuracy of $45.00 \pm 3.13\%$ across all subjects. In the BCI Competition DB [44], the 5-class classification accuracy using our proposed pretraining framework shows an average accuracy of $48.10 \pm 3.68\%$ across all subjects.

We compared four state-of-the-art methods. In addition, to determine how effective deep metric learning is compared to general deep learning, we classified the base class using the cross-entropy loss based on the proposed CNN architecture. Finally, we also compared the classification performance using an SVM instead of the k-NN classifier in the proposed framework. Table III shows the classification accuracy of different methods averaged across all the subjects compared to the proposed method. When we compared the classification performance of the proposed pretraining framework and conventional methods, there were significant differences between the accuracies of the different methods ($p < 0.001$). The proposed method outperformed the state-of-the-art methods by 23.53% to 27.54% in Coretto DB [26]. Similarly, in the BCI Competition DB [44], we achieved higher performance than state-of-the-art methods, ranging from 10.86% to 28.31%. We also showed significantly higher performance even when

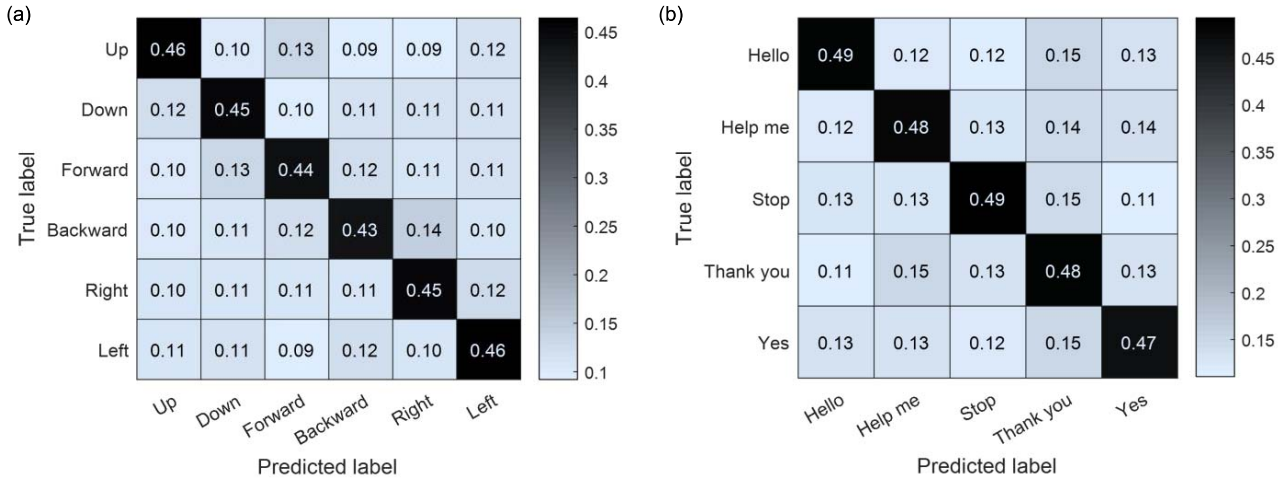

Fig. 2. Average confusion matrix of baseline classification across all subjects without incremental learning. (a) 6-class in Coretto DB [26] and (b) 5-class in BCI Competition DB [44] during imagined speech.

TABLE III
COMPARISON OF ACCURACY RATE WITHOUT INCREMENTAL LEARNING

Dataset	Method	Acc. ± Std.	p-value
Coretto [26]	Coretto et al. [26]	17.46 ± 0.75	<0.01
	Cooney et al. [33]	18.89 ± 1.41	<0.01
	Schirrmeister et al. [35]	19.81 ± 2.10	<0.01
	García-Salinas et al. [10]	21.47 ± 1.99	<0.01
	CNN	22.01 ± 2.33	<0.01
	Proposed method+SVM	38.67 ± 3.87	<0.01
	Proposed method	45.00 ± 3.13	-
BCI Competition [44]	Coretto et al. [26]	37.24 ± 6.24	<0.01
	Cooney et al. [33]	19.79 ± 1.45	<0.01
	Schirrmeister et al. [35]	20.06 ± 1.05	<0.01
	García-Salinas et al. [10]	23.26 ± 1.66	<0.01
	CNN	23.54 ± 2.00	<0.01
	Proposed method+SVM	37.33 ± 2.80	<0.01
	Proposed method	48.10 ± 3.68	-

using the proposed methods with SVM and CNN, which are part of the proposed method, on both datasets. As a result, the proposed method in both datasets statistically performed better than the other methods. This finding indicates that the proposed framework is suitable for classifying imagined speech.

To visualize the levels of similarity between features in each class, we depicted the multidimensional scaling (MDS) plot and performed an agglomerative hierarchical cluster analysis in both datasets. Fig. 3(a) illustrates the embeddings of the data distribution obtained through the proposed feature extraction method. If features of each class were far enough away to be classified, this means that the EEG features through the Siamese network are well extracted. The raw EEG data were extracted from the first-order feature through the instantaneous frequency and spectral entropy, and the second-order feature was extracted through the Siamese neural network. The different classes are clearly distinguished from each other in the MDS plot. Fig. 3(b) shows that features of each class are paired to form a cluster based on similarity. On the two datasets, each class was not hierarchically close to a particular class, all classes were at similar distances. This means that there are no visibly particularly close or distant classes, and each class is distributed at a certain level of distance from each other so that they can be distinguished from each other. In this regard, this could be evidence to support that the Siamese network has extracted distinguishable features in each class during the imaged speech.

In addition, we compared the performance using raw EEG signals to investigate whether the instantaneous frequency and spectral entropy extract important features of imagined speech. The average classification performances for all the subjects are shown in Supplementary Fig. S2. A higher classification accuracy using the instantaneous frequency and spectral entropy was observed than when using raw EEG signals.

C. Classification of Novel Class

Table IV shows the classification performance in the Coretto DB [26] and BCI Competition DB [44]. The baseline represents the classification accuracy using the base class (step 1). The adaption method shows the classification performance using novel class detection (step 2). The individual accuracy is shown in Supplementary Tables S1 (Coretto DB [26]) and S2 (BCI Competition DB [44]). The performance of incremental learning is slightly lower than the baseline accuracy. However, there were no significant differences in the classification performance in either dataset. Fig. 4 is a confusion matrix using incremental learning. This finding shows that, even with incremental learning, the classification performance is maintained regardless of the adapted class. This result also shows that our proposed method is suitable for incremental learning. It also indicates that fine-tuning is properly learning without breaking the network. Confusion matrices for all the classes can be found in Supplementary Fig. S3.

We compared our method with that of García-Salinas *et al.* [10] for expanding the number of

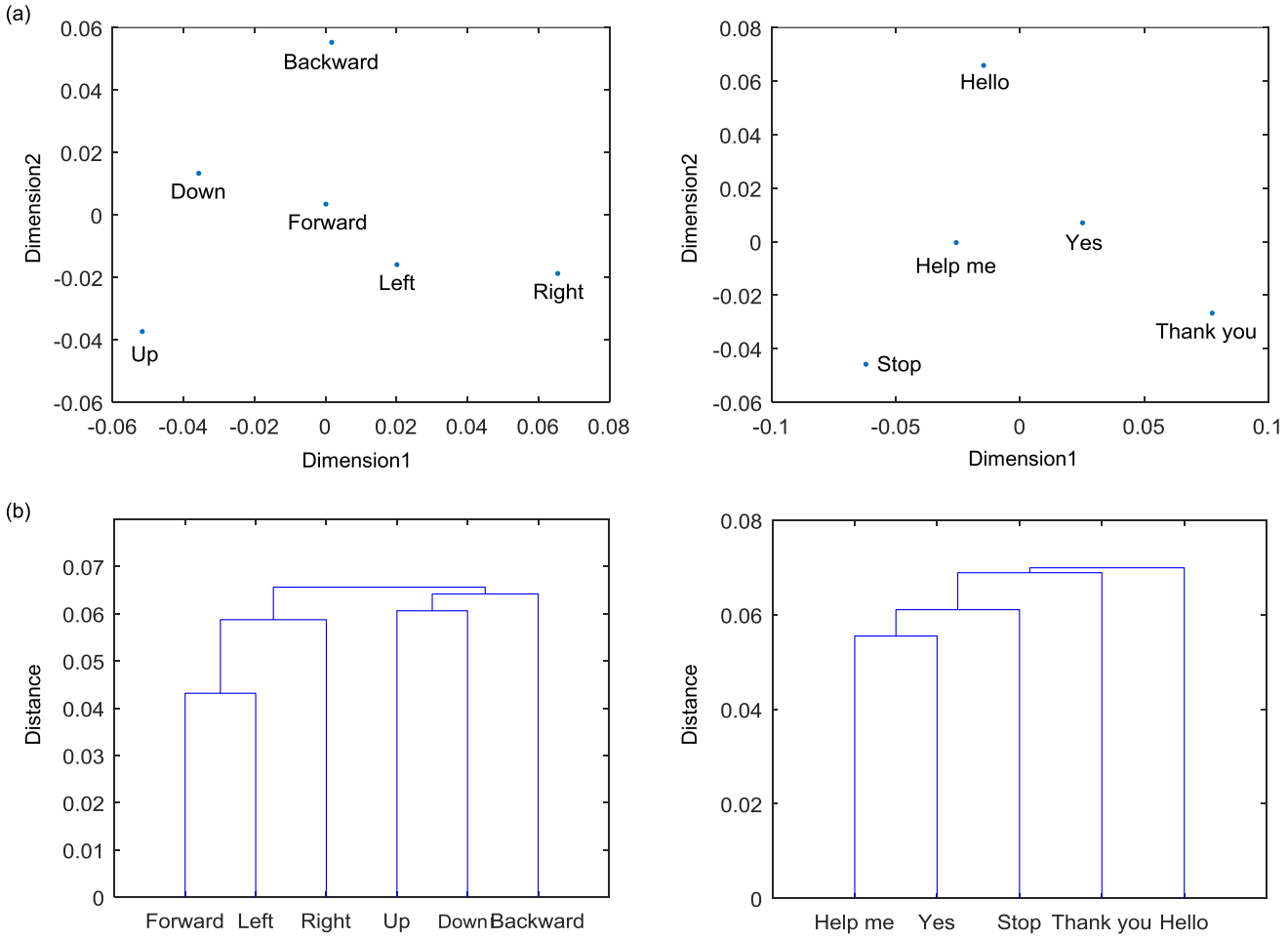

Fig. 3. Similarity among the features in each class extracted through Siamese networks. (a) Multidimensional scaling (MDS) plot. These plots show the Euclidean distances between these classes generated by the Siamese neural network. (b) Dendrogram plot on hierarchical cluster in (left) 6-class of Coretto DB [26] and (right) 5-class of BCI Competition DB [44].

TABLE IV
COMPARISON OF 6-CLASS AND 5-CLASS CLASSIFICATION PERFORMANCE WITHOUT OR WITH INCREMENTAL LEARNING

Dataset	Method	Incrementally learned class	Acc. ± Std.
Coretto [26]	Baseline	-	45.00 ± 3.13
	Incremental learning	Up	43.92 ± 2.52
		Down	44.56 ± 3.10
		Forward	44.63 ± 2.49
		Backward	44.72 ± 2.86
		Right	44.67 ± 2.41
		Left	44.49 ± 2.49
BCI Competition [44]	Baseline	-	48.10 ± 3.68
	Incremental learning	Hello	47.52 ± 3.91
		Help me	47.05 ± 3.48
		Stop	47.14 ± 3.58
		Tank you	46.67 ± 4.19
		Yes	47.24 ± 3.78

classes in imagined speech. The average classification accuracy of all the subjects for each adaptation class is shown in Fig. 5. Our performance was statistically higher for all the classes in both datasets. Even in the paper by García-Salinas *et al.* [10], the different classes of incremental learning greatly affect the performance. In the paper by García-Salinas *et al.* [10], there is a significant difference for all classes. However, the performance difference for our proposed method is not significant for the class to which it is adapted (Table V).

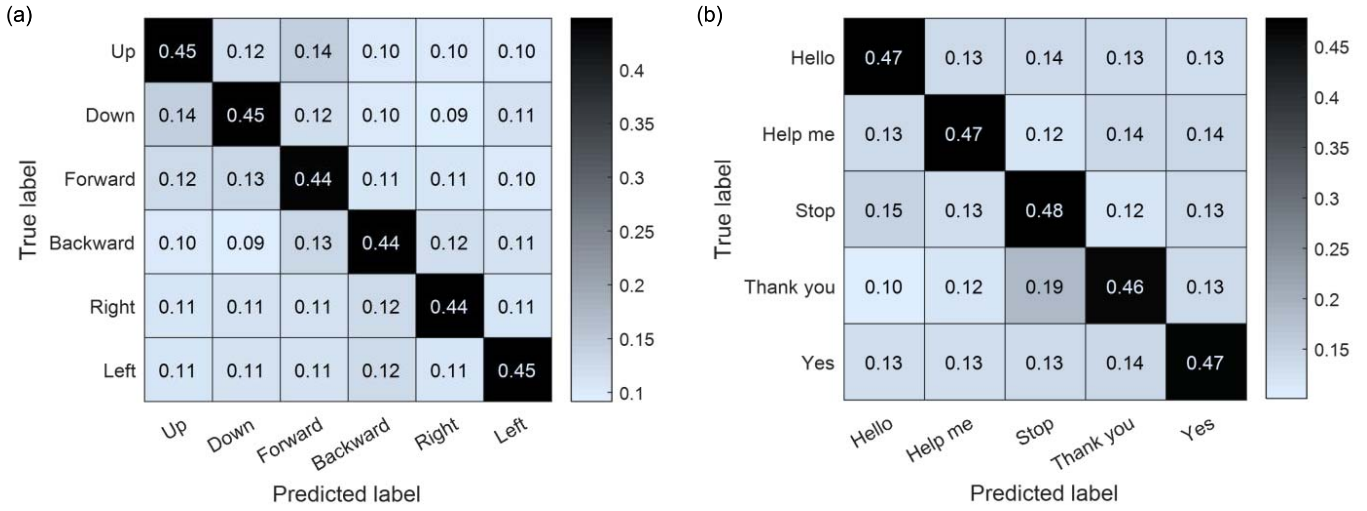

Fig. 4. Averaged confusion matrix of the classification across all subjects with incremental learning. In (a) the Coretto DB [26], 'left' was the novel class, and in (b) the BCI Competition DB [44], 'yes' was the novel class.

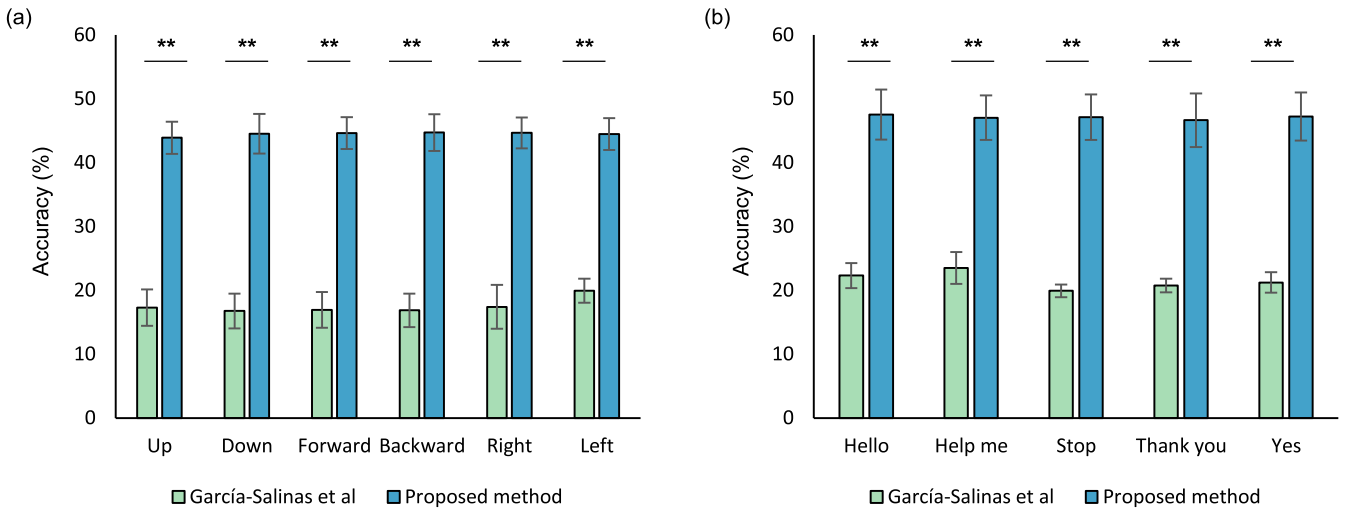

Fig. 5. Comparison of the averaged accuracy between the method used by García-Salinas *et al.* [10] and the proposed method with incremental learning. (a) The 6-class classification performance using the Coretto DB [26] and (b) the 5-class classification performance using the BCI Competition DB [44]. The error bars indicate the standard deviation. ** indicates p-value < 0.01.

TABLE V
STATISTICAL RESULTS BETWEEN CLASSIFICATION PERFORMANCE

Dataset	Method	Acc. ± Std.	p-value
Coretto [26]	García-Salinas et al. [10]	17.54 ± 1.10	0.03
	Proposed method	44.50 ± 0.27	0.97
BCI Competition [44]	García-Salinas et al. [10]	21.54 ± 1.24	<0.01
	Proposed method	47.12 ± 0.28	0.98

We additionally compared the classification performance in the base class before and after incremental learning to explore whether the original capabilities were well preserved (Fig. 6). In other words, this figure indicates the 5-class (Coretto DB [26]) and 4-class (BCI Competition DB [44]) classification accuracy using the base class detection framework excluding the corresponding class. In all the classes, there was no significant difference in the pretraining and incremental learning performance. The results show that incremental learning it does not affect the base class. The proposed method implies that it is a possible method for increasing the number of classes imagined speech to infinity.

V. DISCUSSION

In this study, we proposed a novel framework based on deep metric learning using instantaneous frequency and spectral entropy. Our proposed method significantly outperforms state-of-the-art methods. There was also no performance degradation through incremental learning when the novel class was tested. We evaluated the extensible BCI system in that

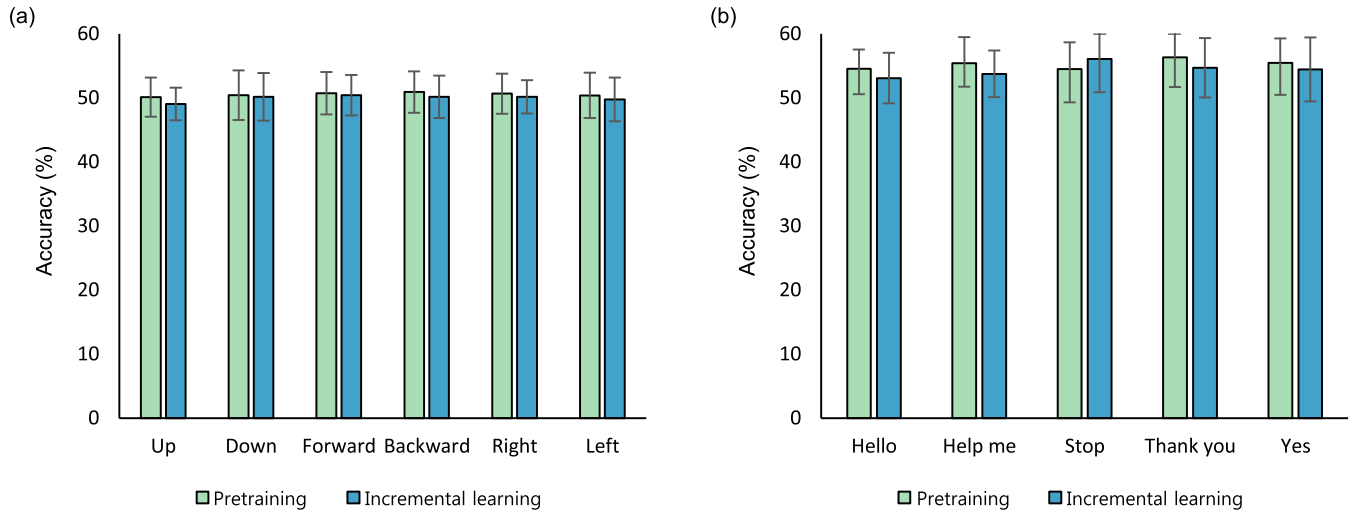

Fig. 6. Comparison of the classification accuracy in the base class before and after incremental learning. This class indicates the incrementally learned class in (a) Coretto DB [26] and (b) BCI Competition DB [44]. The error bars indicate the standard deviation.

it was tested with only a small number of trials using a pretrained classifier. Therefore, the proposed framework shows the possibility of multiclass scalability for BCI communication systems.

We used the instantaneous frequency and spectral entropy inspired by speech signals. During imagined speech, Wernicke's area (the superior temporal gyrus and superior temporal sulcus) and Broca's area (the inferior frontal gyrus) are activated. These regions are essential for speech comprehension and production [46]. In particular, Wernicke's area is involved in both speech recognition and output [47], and neural responses over the superior temporal gyrus are believed to encode the amplitude of the perceived envelope [48]. Therefore, the characteristics of speech signals are considered to be well represented during imagined speech. The classification performance using the proposed features was also higher than that using raw EEG signals. The evidence is that the proposed features based on time-frequency distribution are suitable for classifying imagined speech. It also shows deep metric learning can learn discriminant features from the instantaneous frequency and spectral entropy of EEG signals. In this regard, it seems that the instantaneous frequency and spectral entropy have characteristics over time during imagined speech, and deep metric learning operates as another feature extractor.

In particular, our proposed method outperformed baseline CNN methods. The traditional CNN uses the cross-entropy loss, which is limited to learning the distance between data because it learns the true predicted probability [49]. On the other hand, the contrastive loss used in this study calculates the distance between the embedding data. It learns to embed close to each of the other samples from the same class and to be embedded at a distance larger than the margin samples from different classes [13], [15], [50]. Additionally, our performance using the k-NN classifier was higher than that using an SVM as the classifier. Since the Siamese neural network reduces the dimension by learning based on the distance of the data, the reduced dimension contains information about the distance. Therefore, it is considered that k-NN, which determines classes based on distance than SVM, can classify the data more accurately.

As a result, our method showed an improvement of more than 22.94% (Coretto DB [26]) and 10.86% (BCI Competition DB [44]) over baseline methods. Moreover, the statistical analysis showed significant differences between our proposed method and other methods. The average accuracy across subjects was $45.00 \pm 3.10\%$ (Coretto DB [26]) and $48.10 \pm 3.68\%$ (BCI Competition DB [44]). Therefore, we confirm that the Siamese neural network reduces high-dimensional data (specifically to 8 dimensions) while minimizing the loss of information. Additionally, our Siamese neural network classifies EEG signals from imagined speech without removing extra noise (i.e., removal of electrooculography). EEG signals have a low signal-to-noise ratio [50], which cannot be solved by preprocessing alone. However, our method seems to partially overcome this problem.

Although deep learning requires many trials, it was not easy to acquire many EEG signals. With small EEG trials, the deep learning method with cross-entropy does not learn well and tends to overfit [51]. However, since the Siamese neural network using contrastive loss was originally proposed in one-shot learning, it can be used even with a small amount of data [50]. The Siamese neural network also uses two pairs of samples as input. Therefore, it has the advantage of being able to increase the number of data points by the number of combinations [14]. Our research has validated all possible combinations to demonstrate improvements in classification performance with a small amount of EEG data using a Siamese neural network. In particular, there were no performance differences depending on the novel class. This finding suggests the possibility that users can extend the class to any word. This is also an important factor in imagined speech because

it is difficult to find a specific word for high performance. Therefore, our framework can extend the imagined speech BCI system to any class the user wants with a small number of EEG signals.

VI. Conclusion and Future Works

In this study, we proposed a Siamese neural network framework based on the instantaneous frequency and spectral entropy to classify imagined speech using EEG signals. Our approach consists of (i) training the Siamese neural network using the conditional spectral moments of the time-frequency distribution and (ii) classifying the obtained embedding data. In addition, we proposed a deep learning approach that can learn well without overfitting even when using a small number of trials. Our results showed that the proposed framework has the potential to classify imagined speech. Specifically, the baseline accuracy improved, and the proposed framework classifies the novel class without any performance degradation. Therefore, our proposed framework has shown the possibility of using the imagined speech paradigm as an intuitive BCI for real-world environments.

In future work, we plan to apply other methods. The proposed method is a one-stream method that extracts features through deep learning. However, we modify the two-stream architecture to simultaneously output feature extraction and classification. Additional research is needed to investigate the characteristics of EEG signals during imagined speech. In this regard, we will explore whether the difference in EEG signals is due to pronunciation or meaning when imagining speech. This research would bring us one step closer to designing an intuitive BCI communication system.

References

[1] V. Lawhern, A. Solon, N. Waytowich, S. M. Gordon, C. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, Jul. 2018, Art. no. 056013.

[2] S.-H. Lee, M. Lee, and S.-W. Lee, "Neural decoding of imagined speech and visual imagery as intuitive paradigms for BCI communication," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 12, pp. 2647–2659, Dec. 2020.

[3] L. Wang, X. Liu, Z. Liang, Z. Yang, and X. Hu, "Analysis and classification of hybrid BCI based on motor imagery and speech imagery," *Measurement*, vol. 147, Dec. 2019, Art. no. 106842.

[4] C. M. Wong *et al.*, "Inter- and intra-subject transfer reduces calibration effort for high-speed SSVEP-based BCIs," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 10, pp. 2123–2135, Oct. 2020.

[5] A. Vuckovic, S. Pangaro, and P. Finda, "Unimanual versus bimanual motor imagery classifiers for assistive and rehabilitative brain computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 12, pp. 2407–2415, Dec. 2018.

[6] M.-H. Lee *et al.*, "EEG dataset and OpenBMI toolbox for three BCI paradigms: An investigation into BCI illiteracy," *GigaScience*, vol. 8, no. 5, pp. 1–16, May 2019.

[7] S.-H. Lee, M. Lee, J.-H. Jeong, and S.-W. Lee, "Towards an EEG-based intuitive BCI communication system using imagined speech and visual imagery," in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, Oct. 2019, pp. 4409–4414.

[8] H.-I. Suk and S.-W. Lee, "Subject and class specific frequency bands selection for multiclass motor imagery classification," *Int. J. Imag. Syst. Technol.*, vol. 21, no. 2, pp. 123–130, 2011.

[9] C. Cooney, R. Folli, and D. Coyle, "Neurolinguistics research advancing development of a direct-speech brain-computer interface," *IScience*, vol. 8, pp. 103–125, Oct. 2018.

[10] J. S. García-Salinas, L. Villaseñor-Pineda, C. A. Reyes-García, and A. A. Torres-García, "Transfer learning in imagined speech EEG-based BCIs," *Biomed. Signal Process. Control*, vol. 50, pp. 151–157, Apr. 2019.

[11] E. B. J. Coffey, S. C. Herholz, A. M. P. Chepesiuk, S. Baillet, and R. J. Zatorre, "Cortical contributions to the auditory frequency-following response revealed by MEG," *Nature Commun.*, vol. 7, no. 1, pp. 1–11, Apr. 2016.

[12] H. Watanabe, H. Tanaka, S. Sakti, and S. Nakamura, "Synchronization between overt speech envelope and EEG oscillations during imagined speech," *Neurosci. Res.*, vol. 153, pp. 48–55, Apr. 2020.

[13] M. Kaya and H. C. S. Bilge, "Deep metric learning: A survey," *Symmetry*, vol. 11, no. 9, p. 1066, 2019.

[14] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 1–8.

[15] R. Thiyagarajan, C. Curro, and S. Keene, "A learned embedding space for EEG signal clustering," in *Proc. IEEE Signal Process. Med. Biol. Symp. (SPMB)*, Dec. 2017, pp. 1–4.

[16] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, nos. 3–4, pp. 187–207, Apr. 1999.

[17] J.-L. Shen, J.-W. Hung, and L.-S. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," in *Proc. Int. Conf. Spoken Lang. Process.*, 1998, pp. 1–4.

[18] Y. N. Pan, J. Chen, and X. L. Li, "Spectral entropy: A complementary index for rolling element bearing performance degradation assessment," *Proc. Inst. Mech. Eng., C, J. Mech. Eng. Sci.*, vol. 223, no. 5, pp. 1223–1231, May 2009.

[19] H. Yin, V. Hohmann, and C. Nadeu, "Acoustic features for speech recognition based on gammatone filterbank and instantaneous frequency," *Speech Commun.*, vol. 53, no. 5, pp. 707–715, May 2011.

[20] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "ICaRL: Incremental classifier and representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2001–2010.

[21] F. M. Castro, M. J. Marín-Jiménez, N. Guil, C. Schmid, and K. Alahari, "End-to-end incremental learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 233–248.

[22] K. Shmelkov, C. Schmid, and K. Alahari, "Incremental learning of object detectors without catastrophic forgetting," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3400–3409.

[23] C. S. DaSalla, H. Kambara, M. Sato, and Y. Koike, "Single-trial classification of vowel speech imagery using common spatial patterns," *Neural Netw.*, vol. 22, no. 9, pp. 1334–1339, 2009.

[24] M. Lee, J.-G. Yoon, and S.-W. Lee, "Predicting motor imagery performance from resting-state EEG using dynamic causal modeling," *Frontiers Hum. Neurosci.*, vol. 14, p. 321, Aug. 2020.

[25] S. Shahtalebi, A. Asif, and A. Mohammadi, "Siamese neural networks for EEG-based brain-computer interfaces," 2020, *arXiv:2002.00904*. [Online]. Available: https://arxiv.org/abs/2002.00904

[26] G. A. P. Coretto, I. E. Gareis, and H. L. Rufiner, "Open access database of EEG signals recorded during imagined speech," in *Proc. 12th Int. Symp. Med. Inf. Process. Anal.*, Jan. 2017, Art. no. 1016002.

[27] S.-K. Yeom, S. Fazli, K.-R. Müller, and S.-W. Lee, "An efficient ERP-based brain-computer interface using random set presentation and face familiarity," *PLoS ONE*, vol. 9, no. 11, Nov. 2014, Art. no. e111157.

[28] M. N. I. Qureshi, B. Min, H. Park, D. Cho, W. Choi, and B. Lee, "Multi-class classification of word imagination speech with hybrid connectivity features," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 10, pp. 2168–2177, Oct. 2017.

[29] D. Dash, A. Wisler, P. Ferrari, E. M. Davenport, J. Maldjian, and J. Wang, "MEG sensor selection for neural speech decoding," *IEEE Access*, vol. 8, pp. 182320–182337, 2020.

[30] P. Saha, S. Fels, and M. Abdul-Mageed, "Deep learning the EEG manifold for phonological categorization from active thoughts," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2762–2766.

[31] P. Saha, M. Abdul-Mageed, and S. Fels, "Speak your mind! Towards imagined speech recognition with hierarchical deep learning," 2019, *arXiv:1904.05746*. [Online]. Available: https://arxiv.org/abs/1904.05746

[32] D. Dash, P. Ferrari, and J. Wang, "Decoding imagined and spoken phrases from non-invasive neural (MEG) signals," *Frontiers Neurosci.*, vol. 14, p. 290, Apr. 2020.
[33] C. Cooney, R. Folli, and D. Coyle, "Optimizing layers improves CNN generalization and transfer learning for imagined speech decoding from EEG," in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, Oct. 2019, pp. 1311–1316.
[34] C. Cooney, A. Korik, F. Raffaella, and D. Coyle, "Classification of imagined spoken word-pairs using convolutional neural networks," in *Proc. Graz BCI Conf.*, 2019, pp. 338–343.
[35] R. T. Schirrmeister *et al.*, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.
[36] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. I. Fundamentals," *Proc. IEEE*, vol. 80, no. 4, pp. 520–538, Apr. 1992.
[37] T. Inouye *et al.*, "Quantification of EEG irregularity by use of the entropy of the power spectrum," *Electroencephalogr. Clin. Neurophysiol.*, vol. 79, no. 3, pp. 204–210, 1991.
[38] M. Ali, L. T. Jung, A.-H. Abdel-Aty, M. Y. Abubakar, M. Elhoseny, and I. Ali, "Semantic-K-NN algorithm: An enhanced version of traditional K-NN algorithm," *Expert Syst. Appl.*, vol. 151, Aug. 2020, Art. no. 113374.
[39] R. Palaniappan, K. Sundaraj, and N. U. Ahamed, "Machine learning in lung sound analysis: A systematic review," *Biocybern. Biomed. Eng.*, vol. 33, no. 3, pp. 129–135, 2013.
[40] V. Pestov, "Is the K-NN classifier in high dimensions affected by the curse of dimensionality?" *Comput. Math. Appl.*, vol. 65, no. 10, pp. 1427–1437, May 2013.
[41] A. Gepperth and C. Karaoguz, "A bio-inspired incremental learning architecture for applied perceptual problems," *Cogn. Comput.*, vol. 8, no. 5, pp. 924–934, 2016.
[42] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, Dec. 2017.
[43] C. H. Nguyen, G. K. Karavas, and P. Artemiadis, "Inferring imagined speech using EEG signals: A new approach using Riemannian manifold features," *J. Neural Eng.*, vol. 15, no. 1, Feb. 2018, Art. no. 016002.
[44] *2020 International BCI Competition*. Accessed: Apr. 2020. [Online]. Available: https://osf.io/pq7vb/
[45] D. Muñoz, C. Narváez, C. Cobos, M. Mendoza, and F. Herrera, "Incremental learning model inspired in rehearsal for deep convolutional networks," *Knowl.-Based Syst.*, vol. 208, Nov. 2020, Art. no. 106460.
[46] S. Martin *et al.*, "Word pair classification during imagined speech using direct brain recordings," *Sci. Rep.*, vol. 6, no. 1, p. 25803, May 2016.
[47] S. K. Scott and I. S. Johnsrude, "The neuroanatomical and functional organization of speech perception," *Trends Neurosci.*, vol. 26, no. 2, pp. 100–107, Feb. 2003.
[48] Y. Oganian and E. F. Chang, "A speech envelope landmark for syllable encoding in human superior temporal gyrus," *Sci. Adv.*, vol. 5, no. 11, Nov. 2019, Art. no. eaay6279.
[49] J. Zhang, K. J. Shih, A. Elgammal, A. Tao, and B. Catanzaro, "Graphical contrastive losses for scene graph parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11535–11543.
[50] S. Stober, "Learning discriminative features from electroencephalography recordings by encoding similarity constraints," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 6175–6179.
[51] S. Zhao and F. Rudzicz, "Classifying phonological categories in imagined and articulated speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 992–996.