# Real-Time Deep Neurolinguistic Learning Enhances Noninvasive Neural Language Decoding for Brain–Machine Interaction

Ji-Hoon Jeong, *Associate Member, IEEE*, Jeong-Hyun Cho, Byeong-Hoo Lee, and Seong-Whan Lee, *Fellow, IEEE*

***Abstract*—Electroencephalogram (EEG)-based brain–machine interface (BMI) has been utilized to help patients regain motor function and has recently been validated for its use in healthy people because of its ability to directly decipher human intentions. In particular, neurolinguistic research using EEGs has been investigated as an intuitive and naturalistic communication tool between humans and machines. In this study, the human mind directly decoded the neural languages based on speech imagery using the proposed deep neurolinguistic learning. Through real-time experiments, we evaluated whether BMI-based cooperative tasks between multiple users could be accomplished using a variety of neural languages. We successfully demonstrated a BMI system that allows a variety of scenarios, such as essential activity, collaborative play, and emotional interaction. This outcome presents a novel BMI frontier that can interact at the level of human-like intelligence in real time and extends the boundaries of the communication paradigm.**

***Index Terms*—Brain–computer interface, deep neurolinguistic learning, electroencephalogram (EEG), neural language decoding.**

## I. Introduction

THE NATURAL interaction between humans and machines enables us to convey human thoughts directly in a real-world environment. In particular, brain–machine interfaces (BMIs) reach technological maturity and translate neural activity into meaningful outputs that might drive communication between the human mind and external devices [1], [2], [3]. BMI has been investigated not to indirectly interact with the machine through other external manipulation devices, such as a keyboard and joystick, but rather as a direct interaction method that can decode human intentions and inform machines [4], [5], [6], [7]. Significant research in BMI systems has yielded marked improvements as assisted-living devices for individuals with motor or sensory impairments, such as stroke and amyotrophic lateral sclerosis (ALS) [8]. In the past few years, the field has expanded toward healthy people to support daily life with promising results, as reported in [9] and [10].

A major challenge for this interaction is to naturally ensemble the human mind and machines, decode intuitive human intentions, achieve high accuracy, and communicate in real time [11], [12], [13]. Recently, speech-imagery decoding from brain signals has been investigated as a BMI paradigm. This is because it captures and decodes neural signals corresponding directly to speech production, thereby enabling a naturalistic mode of communication [14], [15]. Recent studies have shown that invasive microelectrode recordings [e.g., electrocardiography (ECoG)] can detect voice activity [16], classify the neural correlates of speech perception [17], and mimic natural dialog [11]. These findings are important steps toward the development of language-based BMI communication that directly decodes speech from recorded neural signals.

Furthermore, for noninvasive scalp electroencephalogram (EEG) signals, a few studies have reported speech imagery decoding. Despite lower-amplitude signals and relatively poor spatial resolution, EEG can provide considerable information by decoding speech imagery over the scalp [18]. It could provide a noninvasive and low-cost means of investigating cortical activity with a high temporal resolution, allowing it to be used for the naturalistic form of the BMI communication system. Recent advances have led to notable success in improving language content decoding directly from EEG. For example, the neural correlates of vowels, consonants, phonemes, syllables, and even words have been classified using advanced decoding algorithms [19], [20]. Despite the promising results achieved to date, EEG-based speech imagery decoding is still a potential field because it has several limitations.

During speech imagery, the neurophysiological response is a complex blend of interaction between the semantic and syntactic factors of given words, thus, neural activities are changed

Manuscript received 21 March 2022; revised 9 July 2022 and 7 September 2022; accepted 29 September 2022. This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korean Government (Artificial Intelligence Innovation Hub) under Grant 2017-0-00451, Grant 2019-0-00079, and Grant 2021-0-02068. This article was recommended by Associate Editor C.-T. Lin. *(Corresponding author: Seong-Whan Lee.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethical Committee of the Institutional Review Board at Korea University under Application No. KUIRB-2019-0143-01.

Ji-Hoon Jeong is with the School of Computer Science, Chungbuk National University, Cheongju 28644, Chungbuk, South Korea (e-mail: jh.jeong@chungbuk.ac.kr).

Jeong-Hyun Cho and Byeong-Hoo Lee are with the Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, South Korea (e-mail: jh_cho@korea.ac.kr; bh_lee@korea.ac.kr).

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

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TCYB.2022.3211694.

Digital Object Identifier 10.1109/TCYB.2022.3211694

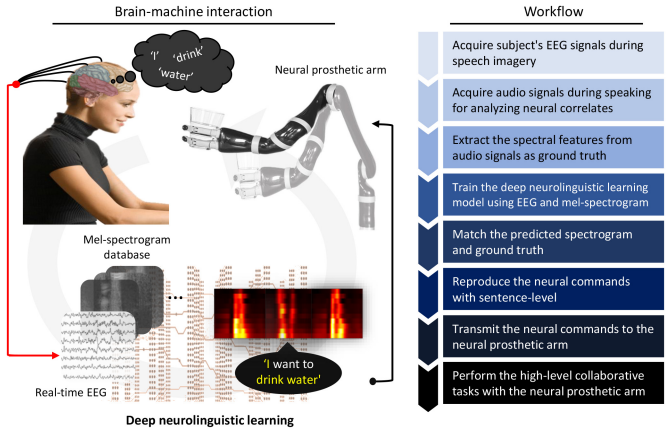


Fig. 1. Overview of deep neurolinguistic learning-based noninvasive neural language decoding for neural prosthetic arm control.

according to each class, participant, and single trial [21]. These phenomena lead to difficulties, such as constraints of the number of classes, deteriorated decoding performance, and nonguaranteed real-time performance. In attempting to decode language content directly to overcome constraint classes, some studies have focused on decoding vowels and consonants [22], syllables [23], phonemes [24], and words [25], [26]. Some studies have demonstrated the classification of auditory sentences/perception from neural correlate activity [27]. Other studies have attempted to use advanced machine learning algorithms to improve the classification of speech imagery units [20], [26], [28]. Compared to many previous incarnations for brain–machine interaction skills [29], speech imagery decoding still exhibits relatively poor performance and has several limitations [21].

Here, we present the possibility of real-time neural language decoding at the sentence level through speech-imagery tasks based on noninvasive brain activity. In this study, neural language is defined as the decoding outcome of a user's speech imagination from brain signals. We provided the most intuitive interaction between humans and neural prosthetic arms through the proposed deep neurolinguistic learning in real time (Fig. 1). To the best of our knowledge, this is the first attempt to collect EEG signals according to the categorization of sentence components into the subject, verb, and object words. Accordingly, we present an intelligent BMI technology that can avoid constraints on the number of classes using deep neurolinguistic learning. To this end, we hypothesized that the neural correlates between the characteristics of the audio and brain signals, and a model that obtained high performance could be proposed by learning it through a deep learning architecture. Hence, in this study, we evaluated the neural language decoding in offline and real-time environments. Based on this, we demonstrated that brain-controlled neural prosthetic arms enable high-performance collaborative tasks with multiple users. This collaboration may function alongside humans and neural prosthetic arms to complete cooperative control using only BMI technology. In this study, we introduced new BMI frontiers that allow users to communicate using a combination of neural language and perform tasks at the level of human-like intelligence with a neural prosthetic arm using only their minds.

## II. Materials and Methods

### A. Participants

Eleven participants (six males and five females) ranging in age from 20 to 34 years were recruited for the experiment. All participants were healthy volunteers and were naïve regarding BMI technology. The volunteers received the experimental protocols, paradigms, and purposes before the experiment. Each participant provided written informed consent in accordance with the Declaration of Helsinki. All experimental protocols were approved by the ethics committee of the Institutional Review Board of Korea University [KUIRB-2019-0143-01].

### B. Experimental Setup and Protocol

Eight words, including subject, verb, and object words, were selected as the most representative elements of a sentence. We selected words that could provide the essential vocabulary for intuitive human–machine interaction, particularly for controlling the neural prosthetic arms. Initially, we collected audio signals of the real words spoken by each participant. The words included "I" and "partner" as subject words, "move," "have," and "drink" as verb words, and "box," "cup," and "phone" as object words. The participants were asked to speak 25 repetitions of each word, as in the speech-imagery sessions.

After recording the audio signals, the participants sat in a comfortable chair and wore a 64-channel EEG actiCap with active Ag/AgCl electrode placement, following the international 10–20 system. The ground and reference electrodes were set as FCz and FPz, respectively. EEG signals were recorded using BrainVision Recorder (BrainProduct GmbH, Germany) with MATLAB 2020a software. The calibration session began after all impedances of the electrodes were less than 10 kΩ. During breaks, the conductive gel was injected into the electrodes using a syringe with a blunt needle.

The EEG recording session was designed to comprise three subsessions according to categorization (first session: subject word; second session: verb word; and third session: object word). The participants conducted speech imagery of the words presented. When the experiment began, visual instructions were provided on a monitor depending on the procedure. Initially, the participants rested comfortably for 3 s. After relaxation, a visual instruction with one of the words was displayed on the monitor as a text sign for 2 s. The participants required a small amount of time (1 s) to prepare, after which they conducted the speech imagery task in four subtrials continuously (Fig. 2). According to each subsession, the participants performed 25 trials per class [30], resulting in 800 trials (25 trials × 4 times × 8 classes) in total. To maintain the physical and mental condition of the participants and, thus, ensure high signal quality, the participants were provided with sufficient breaks (approximately 10–15 min) between each subsession. If they reported an inconvenient position or unstable mental condition (e.g., fatigue), we either adjusted

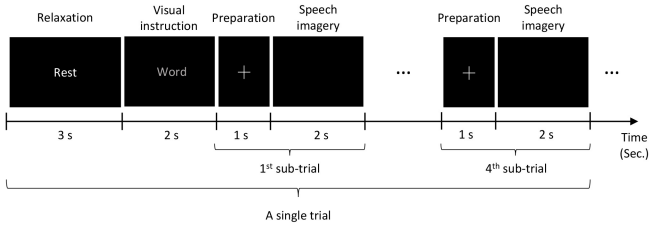


Fig. 2. Experimental setup and protocols in the calibration session. A trial had a duration of 17 s in total. Instructions were provided through visual cues before the participants performed speech imagery. The participants were directed to imagine each of the designated words only once during an imagery period; therefore, only one data point was included in a single subtrial. As a result, four EEG data samples could be obtained via a single trial, which included four subtrials. We performed 50 trials per class and collected a total of 200 data samples per imagined word class.

the experimental environment according to their requests or paused the experiment temporarily [31].

### C. Deep Neurolinguistic Learning

*1) Signal Preprocessing:* The EEG signal was first downsampled from 1000 to 500 Hz, and then a band-pass filter in a range of [30–125] Hz was performed using Hamming-windowed zero-phase finite impulse response (FIR) filters with an optimized order ($N = 50$). We selected a band-pass filter in the range of [30–125] Hz to allow CNN to consider a wider latent space. The EEG data were downsampled from 1000 to 500 Hz. The spectrogram images were obtained by combining the 64 channels into one. Although speech imagery is mainly related to Wernicke's and Broca areas, the study employed whole EEG channels to evaluate how EEG signals can be dispersed by the scalp and skull.

*2) Spectrogram Embedding:* To apply our developed methodology correctly, a mel-spectrogram that expresses speech signals from EEG signals must be estimated. The mel-spectrogram was transformed with a window size of 1024, a hop size of 256, 1024 points of a Fourier transform, and a sampling rate of 22 050 Hz [32]. To build a spectrogram of ground-truth per word class, the average value of the spectrogram was calculated using 25 samples. We would like to use spectrogram images that were as simplified as possible to train deep learning models based on these mel-spectrograms. Therefore, the simplified images through spectrogram embeddings are generated with the downsampling techniques we create. After extracting the original mel-spectrogram from all participants, the spectrograms were represented in a simplified form for model training. We simplified the spectrograms using feature embedding, which is a downsampling strategy. The embedding of spectrograms involves downsampling and reversing the existing figures to positive notes. This process allowed us to obtain the most representative mel-spectrogram to define the ground truth for each individual. Data were sampled by reducing the resolution of mel-spectrogram, which is raw data, and a low-resolution spectrogram image was generated through this. In addition, the reason for the reversal of the negative values is to make it easier for the model to learn, and the part expressed when voice occurs in the mel-spectrogram is converted into a positive value. The surrounding noise is made as zero as possible so that the model can focus on the part where voice occurs when learning.

In particular, downsampling was applied to the frequency bands and time axes, reducing the mel-spectrogram data to $28 \times 28$. The optimal size we found was $28 \times 28$, and reducing the image size to less than that resulted in poor classification performance; reducing the image size to more than that resulted in the same or no significant change in the performance, although the learning time of the model increased. We empirically obtained the estimated spectrograms using these embedding techniques, which allowed the learning model to perform better in progressing classification because the target label of regression becomes simpler. It is impossible to accurately describe the original speech characteristics using a simplified spectrogram via embeddings. However, this strategy can be regarded as a highly efficient preprocessing technique for data because our model has sufficient resolution to proceed with predictions based on images.

In this embedding process, the deep neurolinguistic learning model compares the similarities of ground-truth images with images of simplified spectrograms estimated from EEG signals as the output via the SSIM algorithms. SSIM with a quality assessment index is based on the calculation of three terms: luminance ($l$), contrast ($c$), and structure ($s$). The final index is a combination of three terms, as shown in

$$\mathrm{SSIM}(x, y) = \left[l(x, y)\right]^{\alpha} \cdot \left[c(x, y)\right]^{\beta} \cdot \left[s(x, y)\right]^{\gamma} \tag{1}$$

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{2}$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{3}$$

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}. \tag{4}$$

In this case, $\mu_x$, $\mu_y$, $\sigma_x$, $\sigma_y$, and $\sigma_{xy}$ are the local mean, standard deviation, and cross covariance of images $x$ and $y$, respectively. We calculated SSIM where $\alpha = \beta = \gamma = 1$. Also, $C_1 = (0.01 \times L)^2$, $C_2 = (0.03 \times L)^2$, and $C_3 = (C_2/2)$, where $L$ is the dynamic range of the specified input image, positive scalar.

*3) Training Model:* We estimated a spectrogram from EEG signals over two stages in the model and finally decoded the words to which the predicted images correspond by comparing them to the ground truth of speech class images representing each word. The learning process is illustrated in Fig. 3.

We first describe the process of recording the EEG signals corresponding to each speech image. The recorded EEG signals were filtered through preprocessing modules and prepared with epoch data, which are traditionally used for EEG analysis. As previously mentioned, we applied the [30–125] Hz range, which is the bandwidth of frequencies mainly used in speech imagery decoding, and used 64 corrected channels for the data analysis [26]. Some of the EEG channels in the frontal lobe, which are highly influenced by eye movement and visual stimulation, were removed from the contaminated factors and transformed into clean EEG signals using infomax independent component analysis (ICA) [31], [33], [34]. The corrected

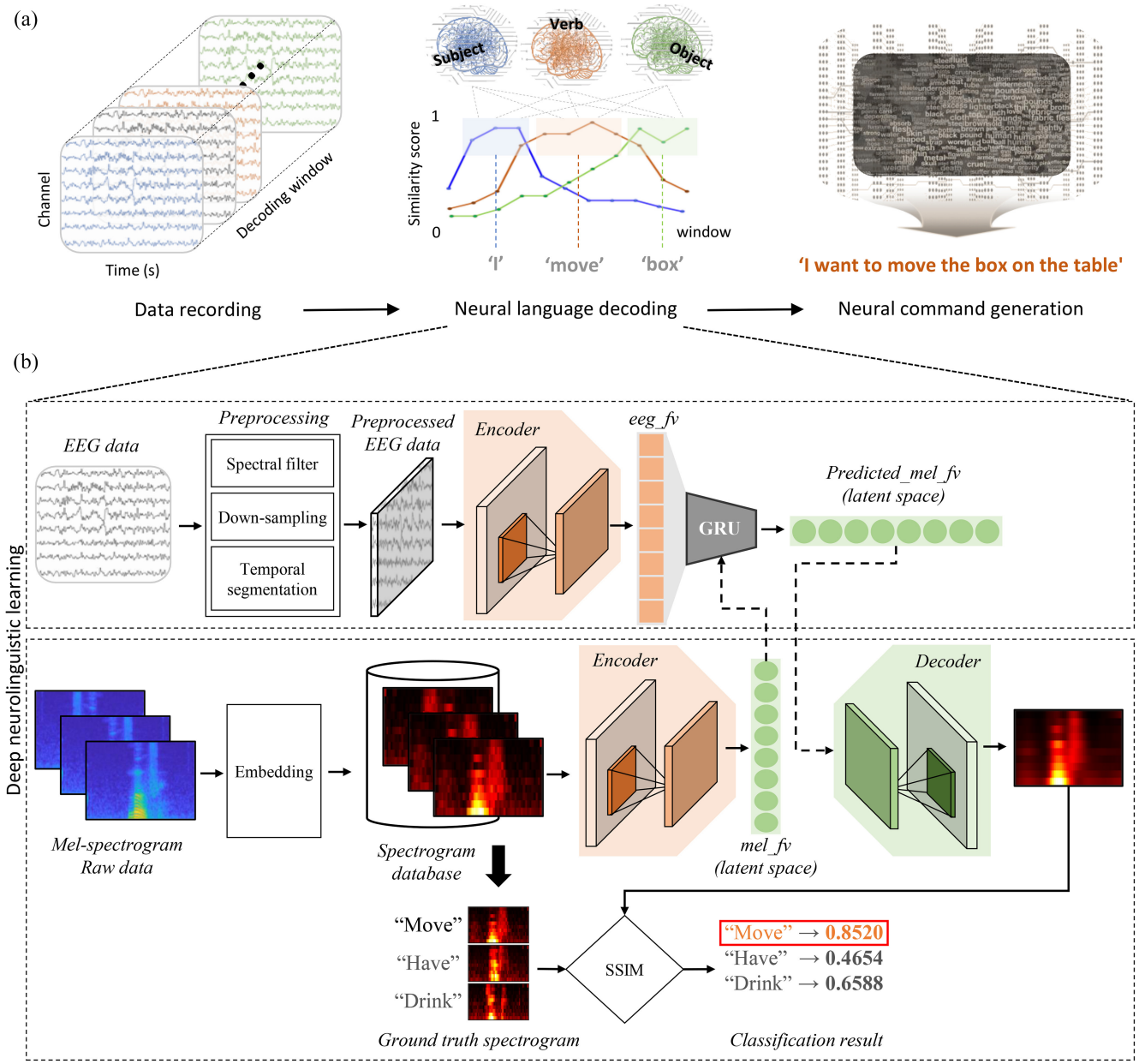


Fig. 3. Deep neurolinguistic learning. (a) Overall flow of the entire system, based on EEG data analysis, categorizes the imagined words and produces meaningful sentences that can eventually allow collaboration with the robot. (b) Detailed illustration of deep neurolinguistic learning. During the training phase, the model receives raw EEG and raw mel-spectrogram as inputs, enabling estimation of the imagined words in the form of a mel-spectrogram through EEG signal analysis. A mel-spectrogram is preprocessed by embedding. Finally, with the SSIM algorithm, the predicted and reconstructed mel-spectrogram is compared to the previously acquired ground-truth spectrogram, each representing a certain class.

EEG signals were normalized using a common average reference (CAR) filter, and preprocessed EEG data divided into lengths of 0–2 s were transferred to the next step and used for learning the model. This preprocessing process of EEG signals, such as band-pass filtering, artifact removal, and normalization, has been equally adapted to real-time experiments. EEG signals are recorded through the filters, embedded in the mel-spectrogram, and leveraged as a label for decoding EEG signals. In the proposed framework, three different CNNs perform the roles of encoder and decoder. The encoder involves convolution layers that can configure the output to be smaller than the input size, whereas the decoder uses deconvolution layers to increase the input size. The deep neurolinguistic learning model first utilizes a CNN to extract meaningful features from EEG data and simplified spectrograms, which progressively simplifies the initial input data using kernels across multiple layers. Each optimized and constructed CNN effectively extracted spatial, temporal, and frequency features from two types of input data (EEG and ground truth) and processed them into data with a length of $1 \times 300$ via the final pooling layer. We transferred $3 \times 1 \times 100$ size features to a module for the next level of the gated recurrent unit (GRU)-based regression [35]. The model learns the applied features to obtain representations of the EEG data and extracts

TABLE I
SPECIFICATIONS OF THE MODEL ARCHITECTURE AND PARAMETERS

| Network | Layer | Type | Parameter | Output size |
|---|---|---|---|---|
| $CNN^{EEG}$ | 1 | Input | - | 1×52×1000 |
| | 2 | Convolution | Filter size: 1×50<br>Stride size: 1×2<br>Feature map: 20 | 20×52×476 |
| | | BatchNorm | - | |
| | 3 | Convolution | Filter size: 1×50<br>Stride size: 1×2<br>Feature map: 40 | 40×52×214 |
| | | BatchNorm | - | |
| | | Activation (ELU) | - | |
| | 4 | Average pooling | Filter size: 1×7<br>Stride size: 1×1 | 40×52×208 |
| | | Dropout | Dropout ratio: 0.6 | |
| | 5 | Depth wise separable convolution | Filter size: 52×1<br>Stride size: 1×1 | 40×1×208 |
| | | BatchNorm | - | |
| | | Activation (ELU) | - | |
| | 6 | Average pooling | Filter size: 1×10<br>Stride size: 1×2 | 40×1×100 |
| | | Dropout | Dropout ratio: 0.6 | |
| | 7 | Fully connected | - | 1×100 |
| | 8 | Fully connected | - | 1×(N:number of classes) |
| | 9 | Softmax | - | Classification output |
| GRU | 1 | Input | - | N×1×100 |
| | 2 | Fully gated unit | Hidden units: 100 | N×100 |
| | | Dropout | Dropout ratio: 0.6 | |
| | 3 | Fully connected | - | 1×(N×100) |
| | 4 | Regression layer | - | 1×100 |

representations of the corresponding spectrogram information as inputs to the regression model simultaneously. The model enables the test phase to estimate data of $1 \times 100$ in size, which can represent the estimated spectrogram with only the input data acquired from the EEG data.

The core part of the CNN consists of a convolutional layer, as shown in Fig. 3. Each convolution layer has several filters to extract the high-level features. Each filter moves successively across the overall spectrogram and extracts the feature values. When the entire spectrogram matches the shape of a particular filter, it has a high value in that part; thus, after the filter traverses the entire spectrogram, it can obtain features only for the part that has a shape similar to that of the filter (Table I). Through these processes, we obtain features that correspond to a particular filter. We used an exponential linear unit (ELU) as an activation function in the convolutional block

$$\mathrm{ELU}(x) = \begin{cases} x, & \text{if} \quad x > 0 \\ a(e^x - 1), & \text{if} \quad x < 0 \end{cases}. \tag{5}$$

A pooling layer is located between the convolution layers and forms a specific area on each feature map, obtained through a filter to derive the largest value within the region. This method is called max pooling, and the derivation of the average value within the corresponding region is called average pooling. This process allowed us to expect more distinctive features that have undergone convolution. The pooling operation also performs downsampling, which maintains the shape, although it can reduce the existing image. A layer of an appropriate depth should be established because the deeper the layers of the CNN, the better the characteristics of the target, and more implications can be extracted. The specifications of the learning model architecture and parameters are listed in Table I.

To complete our decoding framework, we adopted GRU as a prediction model. GRU has the advantage of requiring less learning time and learning with fewer data owing to the small number of parameters [36]. Our model was designed to achieve stable performance under exceptional conditions, in contrast to when there is a sufficient amount of data. The GRU has a simpler structure than long short-term memory (LSTM) [37] to solve the long-term dependency problem and reduce the computation of hidden state updates each time. GRU is a type of recurrent neural network (RNN) framework with gate mechanisms inspired by LSTM. GRU has a structure similar to that of conventional LSTMs and is characterized by a simpler structure, which allows for faster and more efficient learning.

First, in a GRU structure, the reset gate works with the aim of properly initializing the historical information. Using the sigmoid function as an output, the values of zero and one were multiplied by the previously hidden layer. The value of the hidden layer at the previous point in time and the information at the present point can be obtained by multiplying the weight and can be expressed as shown in

$$r^{(t)} = \sigma\left(W_r h^{(t-1)} + U_r x^{(t)}\right). \tag{6}$$

Next, in the GRU, the update gate combines the forget gate and input gate of the LSTM and determines the rate of updating of past and present information. In the update gate, the output $u^{(t)}$ to the sigmoid determines the amount of information at this point, and $(1 - u^{(t)})$ which subtracts the output from one, multiplies the information in the hidden layer at the final moment, which is similar to the roles of the input gate and target gate of the LSTM. This is expressed by (7), as follows:

$$u^{(t)} = \sigma\left(W_u h^{(t-1)} + U_u x^{(t)}\right). \tag{7}$$

The candidate layer is used to calculate the candidate information groups at this time. The key point is to multiply the results of the reset gate rather than use the information of the previously hidden layer. The vertical representation is as follows. In (8), $\tau$ is a tangent hyperbolic, and $*$ denotes the pointwise operation

$$\bar{h}^{(t)} = \tau\left(Wh^{t-1}\right) * \tau^{(t)} + Ux^{(t)}). \tag{8}$$

Finally, the hidden layer combines the results of the updated gate with the candidate result. As previously stated, the result of the sigmoid function determines the amount of information in the final result, and the result of the 1-sigmoid function determines the amount of information at a previous point in time, as summarized in

$$h^{(t)} = \left(1 - u^{(t)}\right) * h^{(t-1)} + u^{(t)} * \bar{h}^{(t)}. \tag{9}$$

In the next step, we provided a feature vector of the estimated spectrogram via the GRU model as an input to the CNN, called the training phase. This network is a CNN design in which weights are trained in the process of extracting feature vectors from the ground truth and is applied as a transpose CNN (with a reversed CNN architecture) to allow the feature vectors to be extended back to the original image data in the

**Algorithm 1:** Deep Neurolinguistic Learning

**Input**:
- EEG data: $X = \{x_i\}_{i=1}^{D}$, $\{x_i\} \in \mathbb{R}^{C\times T}$,
- Mel-spectrogram: $Y = \{y_i\}^{G}$, $\{y_i\} \in \mathbb{R}^{C\times T}$
- $D$ : Total number of trials
- $G$ : Corresponding class
- $C$ : Channels
- $T$ : Time points

**Output**:
- Predicted words: $W = \{w_1, w_2,\ldots, w_k\}$

**for** $k = 1$ ***to*** $K$ **do**
  $X_k = \{x_1, x_2,\ldots, x_i\}$;
  $Y_k = \{y_1, y_2,\ldots, y_i\}$;
  **switch** *Training CNN* **do**
    Train $\text{CNN}_k^{EEG}$ encoder;
    Training the $\text{CNN}_k^{MEL}$ encoder and decoder;
    Create $eeg\_fv_k$ and $mel\_fv_k$;
    Save $eeg\_fv = \{eeg\_fv_1, eeg\_fv_2,\ldots, eeg\_fv_k\}$;
    Save $mel\_fv = \{mel\_fv_1, mel\_fv_2,\ldots, mel\_fv_k\}$;
  **endsw**
  **switch** *Training GRU* **do**
    Train the GRU;
    Create $\overline{mel\_fv}_k$;
  **endsw**
**end**
**if** $k = K$ **then**
  Run $\text{CNN}_k^{MEL}$ decoder using $\overline{mel\_fv}_k$;
  Create $\bar{y}$;
**end**
**if** $k = K$ **then**
  Recall $Y_k = \{y_1, y_2,\ldots, y_i\}$;
  **for** $i = I$ **do**
    Calculate the similarity scores between $\bar{y}$ and $y_i$;
    Rank the similarity scores with the corresponding $y_i$.
    $w_k = i^{th}$ word;
  **end**
**end**
$W = \{w_1, w_2,\ldots, w_k\}$;

**Algorithm 2:** R-SGM

**Function**: Generation of neural language as a sentence-level form

**Input**:
- Predicted words: $W = \{w_1, w_2,\ldots, w_k\}$
- $w_k$ : A predicted word as one of the components of a sentence, obtained through a deep neurolinguistic learning model
- $K$ : The number of $w_k$

**Output**:
- Sentence generated as the neural command
- $\hat{S} = \{$"*Predicted sentence*"$\}$

**for** $k = 1$ ***to*** $K$ **do**
  $W = \{w_1, w_2,\ldots, w_k\}$
  **if** $k = 1$ **then**
    **switch** $w_1$ **do**
      **case** *subject word*
        subj $= w_1$; S$\{k\}$ = subj;
      **endsw**
    **endsw**
  **end**
  **else if** $k = 2$ **then**
    **switch** $w_2$ **do**
      **case** *verb word*
        verb $= w_2$; S$\{k\}$ = verb;
      **endsw**
    **endsw**
  **end**
  **else if** $k = 3$ **then**
    **switch** $w_3$ **do**
      **case** *object word*
        obj $= w_3$; S$\{k\}$ = Obj;
      **endsw**
    **endsw**
  **end**
**end**
Call the pre-assigned sentence example database.
**for** $n = 1$ ***to*** $N$ **do**
  **for** $k = 1$ ***to*** $K$ **do**
    By comparing S$\{k\}$ with components of a sentence
    **if** *Mismatch between S{k} and sentence example* **then**
      return;
    **end**
    Update the pre-assigned sentence including S$\{k\}$ into $\hat{S}$
  **end**
**end**
**if** *Composition requirements of the sentence are met* **then**
  display($\hat{S}$);
**end**

form of 2-D metrics. The image of the spectrogram, which is estimated by EEG decoding, was compared with the prepared ground truth using the SSIM algorithm to determine which class of images is most similar to the ground-truth image representing the corresponding class. Through this process, a deep neurolinguistic model can be used to determine the final prediction. The overall training procedure of the model is summarized in Algorithm 1.

Using the proposed method, we inferred which word the participant imagined from the EEG. Simultaneously, meaningful commands can be generated to interact with the neural prosthetic arm using a combination of these predicted words. To this end, for natural interaction, we generated the neural command as a sentence form consisting of words. In this study, we devised a machine learning model for generating neural commands, which are the final commands that are passed to the neural prosthetic arms generated by combining words classified through BMI decoding, with a simple rule-based sentence generation model (R-SGM). As depicted in Algorithm 2, R-SGM is used to target the most stochastically high sentence composition units, sequencing them in the order of subject words, verb words, and object words. R-SGM is used to reobtain misdecoded results. R-SGM-based neural command generation leverages the sequence of sentence constructs and components that the model learns in advance. Thus, we pre-established the sentence example databases and added them when a similar sequence of sentence components was entered (Algorithm 2). Neural commands at the sentence level can not only affect the most intuitive instruction in natural collaborative control with brain signal-based robots but also increase the degree of freedom of neural commands to large scalability. Because communication in our language involves infinite degrees of freedom, it can be achieved using only EEG.

### D. System Evaluation

*1) Real-Time Decoding Strategy:* To conduct real-time experiments that can decode words imagined by users and organize them into sentences, we established an appropriate experimental environment. The participants imagined a combination of three words, that is, a subject, a verb, and an object, in a single sentence. In this way, the combination of words imagined by the participant can be accurately decoded and

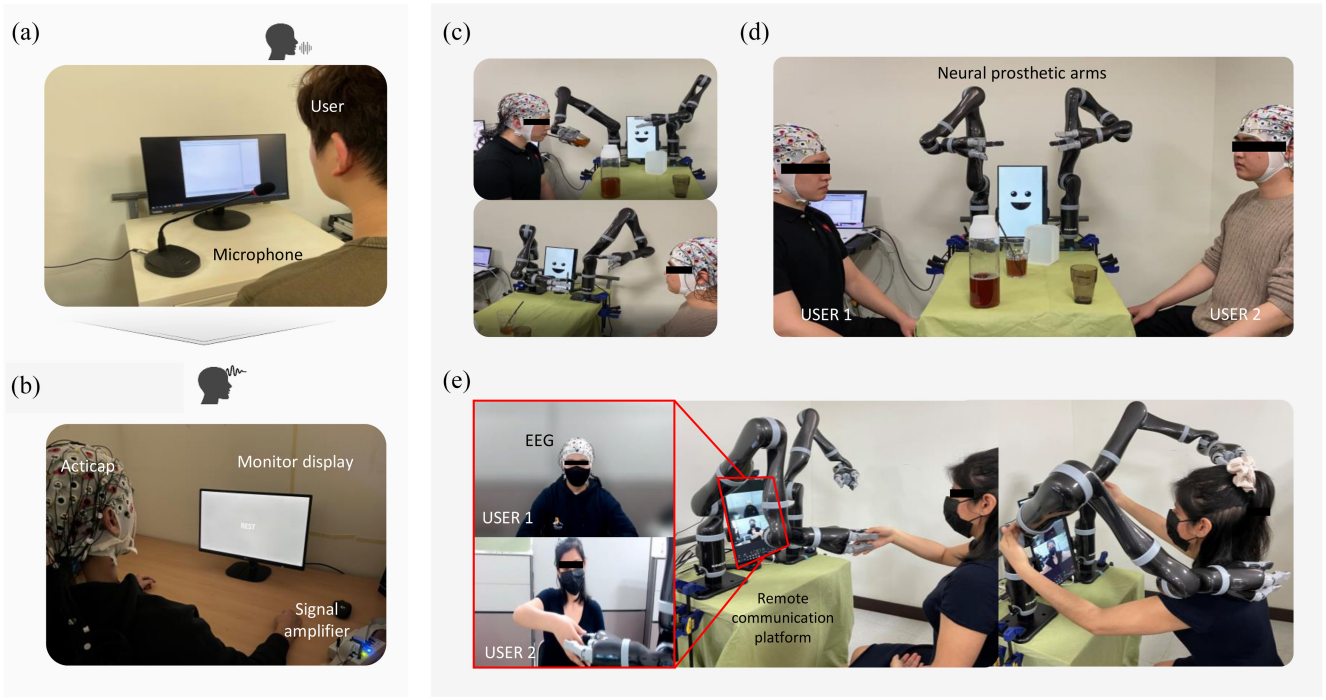


Fig. 4. Calibration experiments and evaluation environments. (a) Process of recording the participants' audio signals to create a mel-spectrogram for the ground truth. (b) Calibration session for acquiring EEG signals while the participants imagine speech using the preassigned words. (c)–(e) Environments of the real-time experiment according to the BMI scenarios for [essential activity (Task I), collaborative play (Tasks II–IV), and emotional interaction (Task V)].

output into one sentence with meaning through the application of a natural language processing model. For example, the previously decoded subject, verb, and object words, I, move, and box are converted into sentences that can convey the clear meaning, "I want to move the box on the table" through the R-SGM, allowing the neural prosthetic arm to recognize the human's thoughts and perform the appropriate actions. In total, 21 subdata segments were obtained by dividing the vertical axis into each word class and the horizontal axis into 2-s decoding windows with 4-s-long data. In a real-time experiment, the participant imagined a sentence using three words for 4 s, that is, a single trial. Specifically, the submodel we developed contained a total of three trials, each with a specialized classification performance for the imagining of each subject, verb, and object word. We guided participants to imagine speaking sentences in a logical manner in the order of subject, verb, and object, and we classified these obtained data with SSIM-based similarity scores, as shown in the submodel. The subject decoding submodel is classified with high similarity in the front segment, whereas the verb segment shows high similarity in the subsequent sequence. Finally, the object decoding submodel outputs a high similarity at the back end of the segmented data. The duration of real-time processing statistically took an average of 4.450 s to decode a single neural language (e.g., a word), and an average of 8.645 s to decode three neural languages to generate neural commands in sentence form in total.

*2) Design of High-Level Collaborative Scenarios:* Finally, we evaluated the entire system by designing real-time BMI scenarios with collaborative tasks using neural prosthetic arms (Fig. 4). We designed various high-level cooperative scenarios comprising five tasks each. The tasks comprised an essential activity (Task I), collaborative play (Tasks II–IV), and emotional interaction (Task V) [Fig. 4(c)–(e)]. The five different tasks were as follows: Task I (drinking juice by the user himself or herself), Task II (providing juice to the partner), Task III (drinking juice with the partner's help), Task IV (delivering a phone in a box to the partner), and Task V (expressing the user's emotions to the partner). In this study, in both a single user and a multiuser environment, the possibility of controlling neural prosthetic arms was validated through neural language decoding, such as drinking water, moving, and bringing objects close to each other. We define high-level cooperative scenarios as a combination of several individual upper-extremity movements in which users perform meaningful actions in real life using a neural prosthetic arm. For example, the simple extension of a neural prosthetic arm and the use of a neural prosthetic arm to hold an object is related to individual upper-extremity movements. In contrast, combining these movements to drink a cup of water or shake hands and hug a partner by using a neural prosthetic arm is a high-level scenario. As depicted in Fig. 4(c), each participant sits in a comfortable position in front of the neural prosthetic arms and engages in a scenario in which the participants alternate their roles. Because decoding a subject word referred to in sentence components is now achievable, we demonstrated this technique for neural prosthetic arm control for the other partner in existing user-centric controllable BMI technologies.

In addition, emotional words, such as "hug," "shake," and "greet," which could share the same emotions in an environment in which motor-disabled patients are assumed, were newly created. For example, in the pandemic environment

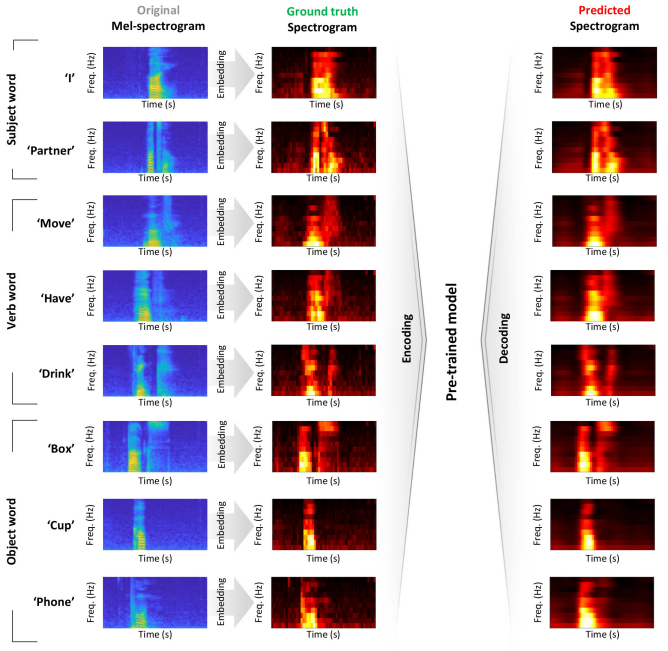


Fig. 5. Example of spectrograms during speech imagery according to each word. Each spectrogram, which includes spectral information, was converted from the original mel-spectrogram to an image simplified by the embedding procedure and used as the ground-truth spectrogram. The learning model was trained on the ground truth according to each word and could predict the spectrogram forms.

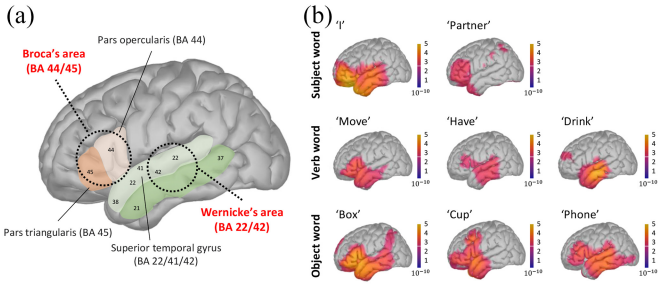


Fig. 6. Neurophysiological characteristic representation during speech imagery according to each word. (a) Diagram depicts brain regions typically associated with the language function in the brain, such as BA44 and BA45 (Broca's area) and Wernicke's area (BA22 and BA42). (b) Spatial representation uses a source imaging technique to activate brain regions statistically, while a participant performs each task. The yellow-colored distribution indicates $p$-values less than 0.001, and the red-colored distribution indicates $p$-values less than 0.005.

when we prefer to communicate in a non-face-to-face manner, the neural prosthetic arms enabled us to express another person's thoughts and deliver their emotions on their behalf in this study. This is a scenario configuration suitable for the current limited context and confirms the feasibility of future BMI technologies being used appropriately in these situations.

## III. Experimental Results

We applied percent valid correct (PVC) as an evaluation indicator [38] due to interaction between brain and robot. If the probability of the classification model is lower than the threshold, it is determined that the brain signals are invalid and are not used for evaluation. In that case, the neural prosthetic arm is not controlled. A deep neurolinguistic learning model was trained using mel-spectrograms extracted from audio signals as the ground truth and preprocessed EEG data. The mel-spectrogram was computed using the short-time Fourier transform (STFT) method [32], and it was visualized differently depending on the subject, verb, and object words. In this study, we adopted the mel-spectrogram, which can be clearly distinguished among audio signals, as a tool for neuronal decoding. Examples of the original mel-spectrogram and embedded spectrogram are presented as the ground truth, as depicted in Fig. 5. Thus, to adapt to the feature embedding, spectrogram transformation was conducted to achieve a form similar to that of the EEG data. For each class, after embedding, the mel-spectrogram is shown in downsampled form as the ground truth. The learning model was trained on the EEG data and an embedded spectrogram. The results of neural language decoding were estimated using the structural similarity index measure (SSIM) [39] between the ground truth and predicted spectrogram. The predicted spectrograms were clearly reflected in the audio configuration features. For example, when the participants imagined the spoken form of words such as "drink," the spectral features at the "in-" ($i\eta$) time points were not evident. However, at the "-nk" ($\eta k$) time points, spectral features appeared. In addition, the duration features were reflected according to the syllable length of the words. For example, compared with the length of the syllable when cup is spoken and the length of the syllable when partner is spoken, we confirmed that the duration of spectral activation is longer and thicker for the partner case. Fig. 5 shows the averaged representation to show qualitatively whether the patterns of spectrogram per class were similar across all participants. In other words, in the calibration session, the original mel-spectrograms and predicted spectrograms per class were individually composed for each participant.

In addition, we visualized a spatial representation using a source imaging technique to statistically identify activated brain regions while the participant performed each task (Fig. 6). Imagining and speaking a language from a neurophysiological perspective is closely related to two regions of the brain. The main regions in the brain are typically associated with language function, with each of the numbered sections indicating one Brodmann area (BA). BA44 and BA45 (Broca's area) were the most common regions associated with speech imagery production. Wernicke's areas (BA22 and BA42) included in the superior temporal gyrus were also observed in cortical activities [21]. At the front end of this loop lies Broca's area, which is connected to the production of language and speech, or language output. At the other end of the superior posterior temporal lobe lies Wernicke's area, associated with the comprehension, processing, and interpretation of words that are spoken or language inputs [40] [Fig. 6(a)]. The two areas are connected by a large bundle of nerve fibers called the arcuate fasciculus. We first identified the quality of the data obtained from the participants by using source localization analysis to confirm that the degree of spatial activation during actual speaking and speech imagery is significantly relevant [40]. Standardized low-resolution electromagnetic

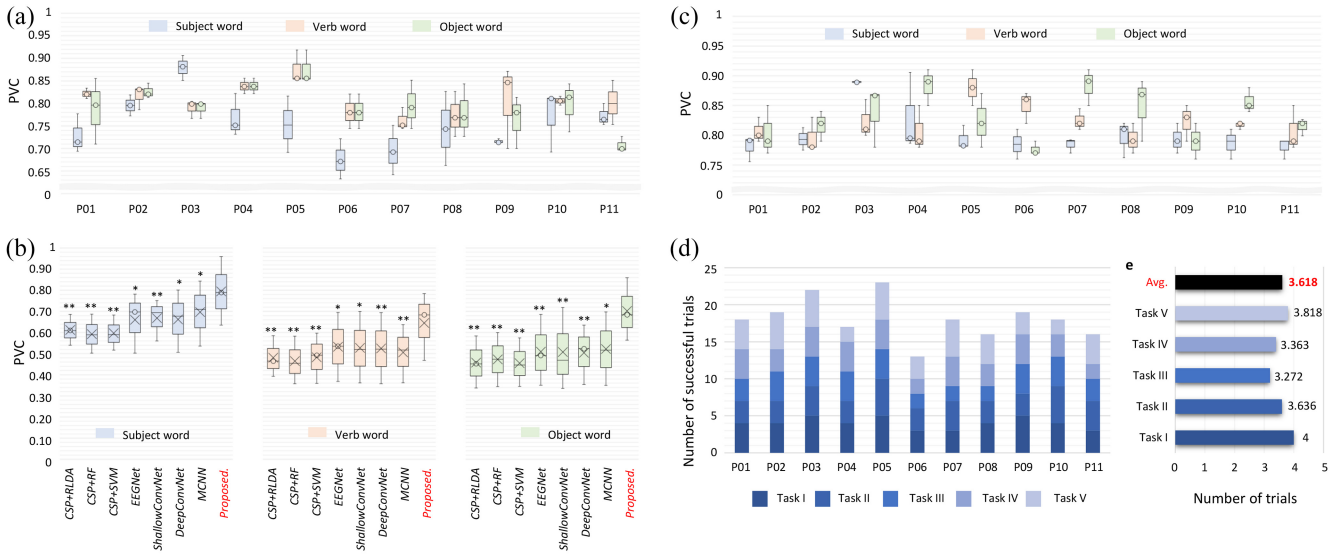


Fig. 7. BMI performance. (a) PVC performance was achieved in the decoding of each word for each participant through a calibration session. (b) Comparison of grand-averaged PVC performance with state-of-the-art methods. The performances showed significant differences according to the subject, verb, and object words. The paired $t$-test was conducted as statistical analysis (*: $p < 0.01$, **: $p < 0.005$). (c) Real-time PVC performance evaluated for each subject, verb, and object word. (d) Number of successful trials in performing high-level collaborative tasks. A total of 25 trials were performed, with five trials for each scenario task. (e) Grand-averaged successful trials per task across all participants.

tomography (sLORETA), which is a current density estimation technique for inverse modeling of brain points, was adapted for visualization [41]. The statistical distribution in the brain regions was visualized by calculating the $p$-values between the baseline and imagery periods. Significant differences were observed mainly in the left cerebral hemisphere, including the lateral sulcus and inferior frontal gyrus for all words. The yellow distribution indicates $p$-values less than 0.001, and the red distribution indicates $p$-values less than 0.005. As shown in Fig. 6(b), we analyzed the source localization obtained during the process of imagining all words from the EEG signals and confirmed that regions of the brain were activated in Broca's area and Wernicke's area related to transitional languages.

Fig. 7(a) shows the PVC performance for generating neural commands using deep neurolinguistic learning in the calibration session. The three-bar plots represent the PVC performance of the sublearning model trained using subject, verb, or object words. In this example, each sublearning model achieved a grand-averaged PVC performance of 0.792 ($\pm$0.058) in the subject words, whereas the verb and object decoding models achieved performances of 0.825 ($\pm$0.037) and 0.811 ($\pm$0.053), respectively, across all participants. The results of the calibration session show that decoding subject-related words consisting of "I" and "partner" is difficult compared to decoding verb and object words. For ten participants (except for P03), the learning model showed lower accuracy in subject words than in verb and object words. This result was also observed in subsequent real-time experiments, and in general, the decoding of verb and object words was relatively similar in the PVC performance. The speech-imagery-based learning model intuitively recalls the imagery of language, although it has not been used in BMI experiments. Thus, even naïve participants can achieve similar performance to experienced people. All the participants we recruited were inexperienced in the use of BMIs, and their performance did not differ significantly from the initial experiment or the subsequent experiment under skilled conditions.

To compare the PVC performance between the baseline methods and the proposed deep neurolinguistic learning, we implemented the machine learning methods presented in [42] and [43] and the deep learning models in [44], [45], and [46]. We evaluated the leave-one-out cross-validation measurements to prevent performance variability. The training epoch was 300, and the test was conducted using weights, which showed the lowest loss after 150 epochs. The AdamW optimizer [47] with 0.001 weight decay and early stopping was applied in deep learning methods. Considering the real-time neural language decoding scenario, PVC was calculated by determining an attempt with a confidence of 0.6 or less as an invalid trial for offline evaluation. For model training, an average of 15 min of running time was spent using the computation resources of an Intel 3.60 Core i7 9700 K CPU with 64 GB of RAM, two NVIDIA TITAN V GPUs, and CUDA/Cudnn. When the trained decoding model was prepared, immediate classification results were obtained for the test data. In a two-step framework, two deep learning models pretrained with training data generated an image from the EEG signals and then classified the mel-spectrogram image (generated from test data). This process required an average of 1.5 s, and the mel-spectrogram image classification required only a brief amount of time, close to the immediate. Under this performance, we conducted real-time online experiments repeatedly without any problem. In addition, to compare the PVC performance between the baseline methods and the proposed method, we conducted a statistical analysis. Initially, we validated the normality and homoskedasticity of each comparative method [CSP+RLDA versus Proposed. in Fig. 7(b)], owing to the small number of samples. The normality for each

baseline method using the Shapiro–Wilk test was conducted for satisfying a null hypothesis (H0), and the assumption of homoscedasticity was also met for each comparative group using Levene's test. Statistical analysis was conducted using a paired $t$-test (*:$p < 0.01$, **:$p < 0.005$). The performances showed significant differences according to the subject, verb, and object words.

Fig. 7(c) shows the results of real-time experiments. All experiments were conducted in a real-world environment, and the learning model was calibrated according to each participant. Each sublearning model was newly trained and tested three times, as represented by the bottom, top, and average of the candles illustrated in the plot. In the real-time experiment, each sublearning model achieved a grand-averaged PVC performance of 0.699 (±0.033) for the subject words, whereas the verb and object decoding models achieved performances of 0.722 (±0.032) and 0.735 (±0.044), respectively.

As mentioned previously, a real-time experiment was conducted after systemically linking the neural prosthetic arms with the BMI system during a calibration session. The success rates of the BMI scenarios with a neural prosthetic arm are shown in Fig. 7(d) and (e). Each high-level collaborative task was performed five times to assess whether the entire scenario was complete. This evaluation involved each participant's ability to perform the BMI tasks, as indicated in Fig. 7(d). Participants achieved an average success rate of 18 trials out of a total of 25 trials. Overall, P03 and P05 each succeeded more than 20 times, showing high scenario performance rates of 88.00% and 92.00%, respectively. P06 exhibited the lowest performance and success rate of approximately 52.00%. Overall, collaborative tasks with simple neural commands, such as Tasks I and V, showed high success rates across all participants, as depicted in Fig. 7(e). In summary, we found that the success rate of each task represented approximately 72.36% of the performance, with 3.618 successes across all participants.

The values of SSIM were presented during a single trial of the real-time experiment, as shown in Fig. 8(a). An example of successful decoding involves P05 who generated a neural command: "I want to move a box on the table." During the 4 s of speech imagery, we observed that the SSIM of the subject part increased first in the order of the components of the sentence, followed by the SSIM of the verb part, and finally that of the object part, resulting in neural language decoding. Specifically, each sublearning model (i.e., subject, verb, and object words) was trained through system calibration beforehand, and decoding was performed simultaneously in real-time according to the step size of the decoding window. While sliding the decoding window over 4 s, the SSIM was calculated, and the word was selected as the final decision by storing the case in which the SSIM was higher than 0.600 three times. When we decoded the neural languages in real time, the threshold of the SSIM was empirically obtained at 0.600. For each sublearning model, 0.875 (4th window in "I"), 0.853 (11th window in "move"), and 0.883 (last window in "box") were recorded as the highest scores.

We also evaluated the spatial distribution to ensure that the participants were fully focused on performing their tasks and

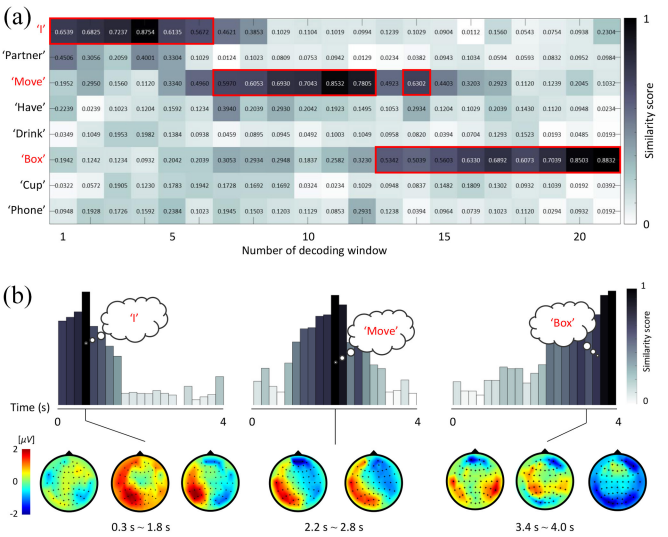


Fig. 8. Decoding process of deep neurolinguistic learning in real time. (a) Each bin represents a segment of real-time EEG data obtained using a decoding window. The similarity score is calculated and compared to the existing ground truth. (b) Bar plots showing the time points in a single 4-s trial for which the subject, verb, and object were classified with a high similarity score. The following scalp plot is intended to depict the correlation between imagined word decoding and neurophysiological evidence. Different colors on the scalp plots represent the average value of the EEG amplitude for each time interval.

to examine whether decoding affected psychological and mental factors that significantly impacted ambient noise or brain signals. Furthermore, during speech imagery tasks, the spatial distribution showed that the left hemisphere, where Broca's and Wernicke's areas were located, exhibited relatively more continuous activation than the right hemisphere [Fig. 8(b)]. In the case of motor imagery tasks, the brain structurally divides body parts, showing strong spatial activation when imagining only or executing tasks [48], [49]. However, because language is not structurally distinct from the brain, it is difficult to distinguish words solely based on the spatial distribution of each component.

## IV. Discussion

The current findings significantly expand previous studies, showing that the neural languages can be decoded in real-time at the noninvasive EEG recording level. We used a novel BMI paradigm called speech imagery, which mimics natural conversations. In state-of-the-art BMI studies, investigators have attempted to provide an intuitive communication pathway between the brain and machines in real-world environments. Recent studies have focused on the intuitiveness of neural commands along with the application of BMI for deep learning technology [50]. This intuitiveness begins by matching the motions of the robot with the neural commands from the human imagination. The motor imagery decoding strategy has advanced from simple body-part imagery (e.g., left hand, right hand, and foot) to intuitive motion imagery, such as arm reaching and hand grasping [31], [51], [52]. Recently, speech imagery, which transcends the limits of the class available as neural commands and decodes speech into brain signals, has

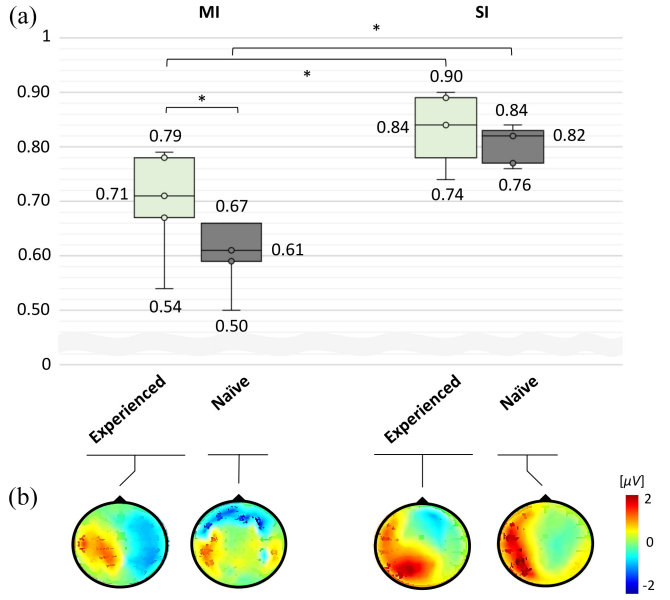


Fig. 9. Comparison of PVC performances between representative BMI paradigms across experienced and naïve participants. (a) Result of measuring the success rate based on the PVC scores when performing a particular task (b) Spatial distributions achieved the average amplitude value of the EEG signals during the tasks (*: $p$-value $< 0.005$).

received attention. The speech imagery paradigm improves the accessibility of BMI for speech-impaired patients or even for healthy people, and demonstrates the feasibility of decoding imagined words from the user's brain signals. Speech-imagery decoding was the first successful method for invasive brain signals [11], [53], [54]. Recently, it has been shown that the speech imagery of simple words can be decoded even from noninvasive EEG signals [21], [26]. Based on these state-of-the-art BMI technologies, in this study, the components of sentences (subjects, verbs, objects, etc.) were decoded rather than simple word decoding, to increase the degree of freedom for users to freely generate practical neural commands. This is the first attempt to adapt the natural communication mechanism to a BMI system that generates neural commands by combining the components of sentences with a variety of subject, verb, and object words at the brain signal level.

Furthermore, through additional experiments that achieved higher decoding performance, we found that speech imagery can be used conveniently for naïve users compared with such as conventional motor imagery (Fig. 9). We recruited five experienced participants and five naïve participants for a simple comparative experiment. The motor imagery paradigm, which is currently used in noninvasive BMI paradigms, and the speech imagery paradigm were adopted in this study. When performing a particular task, we evaluated the decoding performance with respect to the simple classes required (e.g., drinking water, picking up a phone, or moving a box from position $A$ to $B$). First, the difference in decoding performance between experienced and naïve users was found to be approximately 10% within MI. We observed that the difference was statistically significant ($p < 0.005$), as shown in Fig. 9. In contrast, for speech imagery, the physical performance difference between experienced and naïve users were approximately 2%, and no statistically significant difference was observed ($p > 0.01$). In other words, motor imagery could present some difficulties for first-time BMI users, and continued BMI training ensured that it would be readily available in real-world scenarios. However, speech imagery is highly capable of simple word detection. In particular, no significant difference between experienced and naïve participants was observed; thus, we could confirm that participants who first encountered BMI could easily perform the task, which is a significant contribution to the commercialization of BMI technology. Furthermore, even within the same experience, speech imagery was shown to be better decoded, and naïve users found the intuitive paradigm more comfortable when conducting their BMI tasks. However, the BCI illiteracy problem still existed when recruiting participants. Although the proposed learning methods could not cover the BCI illiteracy problem, we will develop learning methods that can be used by a variety of age groups and genders. Hence, we plan to establish training strategy ideas for BCI illiterates before the data-recording session and apply few-shot learning techniques that show reliable performance with only a small amount of learning data algorithmically.

In addition, the results demonstrated that a few participants showed similar performance trends according to a training session and real-time experiments, whereas others showed completely different performance trends [Fig. 7(a) and (c)]. In particular, because of analyzing the tendency of decoding PVC performance for each element of a sentence (i.e., subject words, verb words, and object words), the participants in P01, P03, P07, and P09 showed similar performance tendencies in both the training session and real-time experiments. For example, the PVC values for participant P03 were as high as 0.8813 for subject words, 0.7988 for object words, and 0.7976 for verb words in the training session. In the real-time experiments, PO3 also showed a performance tendency with PVC values in the order of 0.7888 for subject words, 0.7667 for object words, and 0.7100 for verb words. Furthermore, the other participants could not form a certain tendency pattern; therefore, the performance trends with PVC values according to the training session and real-time experiments were different. Participants with different performance tendencies between the training session and real-time experiment were expected to be induced with a few concentration variances owing to the self-directed and continuous decoding (i.e., speech imagery) in real-time experiments, in contrast to training sessions that provide task instruction directly. In the asynchronous BMI experiment, most users wondered if the intended task was accurately decoded. In particular, for asynchronous real-time experiments, the experimental protocol does not simply perform a single imagery task without feedback, although it is asked to conduct each imaginary task (i.e., the components of a sentence) continuously. Based on this, neural commands at the sentence-level form are created. Therefore, the difference in performance tendency between sessions was revealed because the participants themselves generated neural commands without feedback after a single task, which even affected the decoding performance. Although asynchronous BCI in real-time should finally be pursued, it seems that sufficient training

with synchronous paradigms will be required to maintain constant decoding performance. Therefore, in real-time experiments, the experimental protocols are separated into training sessions, synchronous training sessions, and asynchronous real-time experiments. Hence, we expect that the participants can maintain a decoding performance tendency between offline and real-time experiments.

In the BMI literature, although many EEG decodings are being conducted based on the convolutional neural network (CNN) [50], it has not yet been reported to achieve high real-time performance. Because of the sensitivity and nonstationary characteristics of EEG, which change with each session, it may not be trained according to the characteristics of the previously trained input. To compensate for the weaknesses of signal decoding using CNNs, the proposed deep neurolinguistic learning is designed to train two encoders together and perform a final prediction with one decoder (Fig. 3). The parameters and configuration of both encoders and the decoder are described in Table I. Each encoder was trained using EEG signals and audio signal features (mel-spectrogram) by adapting the principle of meta-learning [55]. Owing to training the mel-spectrogram together, it was able to contribute to the model's consistent class learning from invariance characteristics of the EEG signals each time. In addition, the prediction probability was increased by predicting the mel-spectrogram using a decoder. This architecture has become a training strategy that can dramatically improve the low EEG decoding performance, even in a real-time environment. It was possible to improve the real-time performance and design of the convergence framework by utilizing unclear brain signals and mel-spectrogram characteristics. Furthermore, in terms of scalability for practicality in the future, the framework should be able to enable the decoding of not only the learned sentence component classes but also words from unknown speech imagery classes. We will continue to focus on automatically relearning or adapting each model to the user's state at a particular time to achieve reliable performance.

The results were rigorously validated using real-time experiments. In contrast to the conventional BMI systems, which are used only as a tool for control by a single user, the proposed system is designed for multiple users who can control the neural prosthetic arms for both themselves and each other simultaneously. Therefore, in this study, intuitive communication was achieved by combining neural languages as sentence form, and the feasibility of a communication system through collaborative play with robots and brain signals only between multiple users was demonstrated. To the best of our knowledge, this is the first attempt to design a scenario in which collaborative play and neural commands are generated for another participant through noninvasive EEG signals. Therefore, most of the participants showed a success rate of approximately 3.6 out of 5 attempts per task, demonstrating that it is possible to imagine continuous speech imagery at the sentence level in real-time (Figs. 7(c) and 8). In contrast to the conventional motor imagery paradigm, which requires continuous concentration of exercise and movement, we found that the real-time success rate was high because it was easier to perform speech imagery. However, in the ablation study, the proposed deep neurolinguistic learning showed a limitation in that the distinction between words with similar pronunciations was not completely learned. For example, the actual pronunciation of words, such as "he" and "she" is different from the elements of subject words; however, the learning model cannot differentiate between EEG signals and mel-spectrograms. In the case of a similar syllable length, it differs from each characteristic of mel-spectrogram in the audio signals distinctly between words, to allow the model to be trained as high-level features. Therefore, in this study, words with sentence components in which the syllable characteristics of words are distinguished were selected and processed. At this stage, the limitations are evident, and we will apply them to interpret and extract more accurate audio information from brain signals.

## V. Conclusion

This article describes a new approach aimed at intuitive interactions between humans and neuroprostheses by decoding neural language using noninvasive EEG signals. We demonstrated that a mind-controlled neural prosthetic arm based on deep neurolinguistic learning can collaborate with humans to complete high-level tasks successfully. In this study, speech imagery was used as a BMI paradigm and extended from simple word classification to create neural commands by combining the components of each sentence (e.g., subject, verb, and object words) in real time. Consequently, this strategy enables the use of a variety of neural languages to naturally drive an external device using only EEG signals.

In addition, this is the first case in which BMI technology is applied to multiple users without restrictions on the number of classes owing to the use of language-based decoding. Although BMI technology is still demonstrated in limited laboratory environments owing to the sensitivity and sensing difficulties of brain signals, this finding suggests that the scope and environments of collaboration with humans and machines will be increasingly wider. We believe that this research will contribute to a new frontier in BMI technology.

## Acknowledgment

The authors thank Prof. Klaus-Robert Müller for the critical discussion of the experiments and manuscript. Byung-Hee Kwon, Ji-Hoon Kim, and Sang-Hoon Lee are also acknowledged for preprogramming of the neural prosthetic arm and for creating the recording program.

## References

[1] G. Dornhege, J. D. R. Millán, T. Hinterberger, D. J. McFarland, and K.-R. Müller, *Toward Brain–Computer Interfacing*, vol. 63. Cambridge, MA, USA: MIT Press, 2007.

[2] G. Müller-Putz et al., "Towards noninvasive hybrid brain–computer interfaces: Framework, practice, clinical application, and beyond," *Proc. IEEE*, vol. 103, no. 6, pp. 926–943, Jun. 2015.

[3] U. Chaudhary, N. Birbaumer, and A. Ramos-Murguialday, "Brain–computer interfaces for communication and rehabilitation," *Nat. Rev. Neurol.*, vol. 12, no. 9, pp. 513–525, 2016.

[4] R. Mane, T. Chouhan, and C. Guan, "BCI for stroke rehabilitation: Motor and beyond," *J. Neural Eng.*, vol. 17, no. 4, 2020, Art. no. 41001.

[5] D. Zhang, L. Yao, K. Chen, S. Wang, X. Chang, and Y. Liu, "Making sense of spatio-temporal preserving representations for EEG-based human intention recognition," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3033–3044, Jul. 2020.

[6] N.-S. Kwak and S.-W. Lee, "Error correction regression framework for enhancing the decoding accuracies of ear-EEG brain–computer interfaces," *IEEE Trans. Cybern.*, vol. 50, no. 8, pp. 3654–3667, Aug. 2020.

[7] J.-H. Jeong, N.-S. Kwak, C. Guan, and S.-W. Lee, "Decoding movement-related cortical potentials based on subject-dependent and section-wise spectral filtering," *IEEE Trans. Neural Syst. Rehabil Eng.*, vol. 28, pp. 687–698, 2020.

[8] L. R. Hochberg et al., "Reach and grasp by people with tetraplegia using a neurally controlled robotic arm," *Nature*, vol. 485, no. 7398, pp. 372–375, 2012.

[9] C. I. Penaloza and S. Nishio, "BMI control of a third arm for multitasking," *Sci. Robot.*, vol. 3, no. 20, p. eaat1228, 2018.

[10] J.-H. Jeong, K.-H. Shim, D.-J. Kim, and S.-W. Lee, "Brain-controlled robotic arm system based on multi-directional CNN-BiLSTM network using EEG signals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, pp. 1226–1238, 2020.

[11] D. A. Moses, M. K. Leonard, J. G. Makin, and E. F. Chang, "Real-time decoding of question-and-answer speech dialogue using human cortical activity," *Nat. Commun.*, vol. 10, no. 1, pp. 1–14, 2019.

[12] J. Fumanal-Idocin, Y.-K. Wang, C.-T. Lin, J. Fernández, J. A. Sanz, and H. Bustince, "Motor-imagery-based brain–computer interface using signal derivation and aggregation functions," *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 7944–7955, Aug. 2022.

[13] J.-H. Jeong et al., "2020 international brain–computer interface competition: A review," *Front. Human Neurosci.*, vol. 16, Jul. 2022, Art. no. 898300.

[14] O. Iljina et al., "Neurolinguistic and machine-learning perspectives on direct speech BCIs for restoration of naturalistic communication," *Brain–Comput. Interfaces*, vol. 4, no. 3, pp. 186–199, 2017.

[15] A. Tankus, I. Fried, and S. Shoham, "Structured neuronal encoding and decoding of human speech features," *Nat. Commun.*, vol. 3, no. 1, pp. 1–5, 2012.

[16] V. G. Kanas, I. Mporas, H. L. Benz, K. N. Sgarbas, A. Bezerianos, and N. E. Crone, "Real-time voice activity detection for ECoG-based speech brain machine interfaces," in *Proc. Int. Conf. Digit. Signal Process. (DSP)*, 2014, pp. 862–865.

[17] G. M. Di Liberto, J. A. O'Sullivan, and E. C. Lalor, "Low-frequency cortical entrainment to speech reflects phoneme-level processing," *Curr. Biol.*, vol. 25, no. 19, pp. 2457–2465, 2015.

[18] M. D'Zmura, S. Deng, T. Lappas, S. Thorpe, and R. Srinivasan, "Toward EEG sensing of imagined speech," in *Human-Computer Interaction. New Trends*. Berlin, Germany: Springer-Verlag, 2009, pp. 40–48.

[19] A. B. Kristensen, Y. Subhi, and S. Puthusserypady, "Vocal imagery vs intention: Viability of vocal-based EEG-BCI paradigms," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 8, pp. 1750–1759, Aug. 2020.

[20] P. Saha and S. Fels, "Hierarchical deep feature learning for decoding imagined speech from EEG," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 10019–10020.

[21] 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.

[22] 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, 2017, Art. no. 16002.

[23] S. Deng, R. Srinivasan, T. Lappas, and M. D'Zmura, "EEG classification of imagined syllable rhythm using Hilbert spectrum methods," *J. Neural Eng.*, vol. 7, no. 4, 2010, Art. no. 46006.

[24] A. Jahangiri and F. Sepulveda, "The relative contribution of high-gamma linguistic processing stages of word production, and motor imagery of articulation in class separability of covert speech tasks in EEG data," *J. Med. Syst.*, vol. 43, no. 2, pp. 1–9, 2019.

[25] A. R. Sereshkeh, R. Trott, A. Bricout, and T. Chau, "EEG classification of covert speech using regularized neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process*, vol. 25, no. 12, pp. 2292–2300, Dec. 2017.

[26] 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, pp. 2647–2659, 2020.

[27] S. Cai, P. Sun, T. Schultz, and H. Li, "Low-latency auditory spatial attention detection based on spectro-spatial features from EEG," 2021, *arXiv:2103.03621*.

[28] C. Cooney, A. Korik, R. Folli, and D. Coyle, "Evaluation of hyperparameter optimization in machine and deep learning methods for decoding imagined speech EEG," *Sensors*, vol. 20, no. 16, p. 4629, 2020.

[29] U. Chaudhary, B. Xia, S. Silvoni, L. G. Cohen, and N. Birbaumer, "Brain–computer interface-based communication in the completely locked-in state," *PLoS Biol.*, vol. 15, no. 1, 2017, Art. no. e1002593.

[30] C. Vidaurre, T. Jorajuría, A. Ramos-Murguialday, K.-R. Müller, M. Gómez, and V. V. Nikulin, "Improving motor imagery classification during induced motor perturbations," *J. Neural Eng.*, vol. 18, no. 4, p. 460b1, 2021.

[31] J.-H. Jeong et al., "Multimodal signal dataset for 11 intuitive movement tasks from single upper extremity during multiple recording sessions," *GigaScience*, vol. 9, no. 10, p. giaa098, 2020.

[32] J. Shen et al., "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2018, pp. 4779–4783.

[33] X. Wu, B. Zhou, Z. Lv, and C. Zhang, "To explore the potentials of independent component analysis in brain–computer interface of motor imagery," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 3, pp. 775–787, Mar. 2020.

[34] X. Gu et al., "EEG-based brain–computer interfaces (BCIs): A survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 18, no. 5, pp. 1645–1666, Sep./Oct. 2021.

[35] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," 2014, *arXiv:1409.1259*.

[36] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2342–2350.

[37] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

[38] B. J. Edelman et al., "Noninvasive neuroimaging enhances continuous neural tracking for robotic device control," *Sci. Robot.*, vol. 4, no. 31, p. eaaw6844, 2019.

[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, pp. 600–612, 2004.

[40] C. J. Price, J. T. Crinion, and M. MacSweeney, "A generative model of speech production in Broca's and Wernicke's areas," *Front. Psychol.*, vol. 2, p. 237, Sep. 2011.

[41] M. F. Hnazaee, E. Khachatryan, and M. M. Van Hulle, "Semantic features reveal different networks during word processing: An EEG source localization study," *Front. Human Neurosci.*, vol. 12, p. 503, Dec. 2018.

[42] M. Bentlemsan, E.-T. Zemouri, D. Bouchaffra, B. Yahya-Zoubir, and K. Ferroudji, "Random forest and filter bank common spatial patterns for EEG-based motor imagery classification," in *Proc. 5th Int. Conf. Intell. Syst. Model. Simul.*, 2014, pp. 235–238.

[43] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Process Mag.*, vol. 25, no. 1, pp. 41–56, Dec. 2008.

[44] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, 2018, Art. no. 56013.

[45] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapp.*, vol. 38, no. 11, pp. 5391–5420, 2017.

[46] S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Bencherif, and M. S. Hossain, "Multilevel weighted feature fusion using convolutional neural networks for EEG motor imagery classification," *IEEE Access*, vol. 7, pp. 18940–18950, 2019.

[47] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[48] B. Wang et al., "Common spatial pattern reformulated for regularizations in brain–computer interfaces," *IEEE Trans. Cybern.*, vol. 51, no. 10, pp. 5008–5020, Oct. 2021.

[49] Y. Zhang, C. S. Nam, G. Zhou, J. Jin, X. Wang, and A. Cichocki, "Temporally constrained sparse group spatial patterns for motor imagery BCI," *IEEE Trans. Cybern.*, vol. 49, no. 9, pp. 3322–3332, Sep. 2019.

[50] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: A review," *J. Neural Eng.*, vol. 16, no. 3, 2019, Art. no. 31001.

[51] A. Schwarz, M. K. Höller, J. Pereira, P. Ofner, and G. R. Müller-Putz, "Decoding hand movements from human EEG to control a robotic arm in a simulation environment," *J. Neural Eng.*, vol. 17, no. 3, 2020, Art. no. 36010.

[52] J.-H. Cho, J.-H. Jeong, and S.-W. Lee, "Neurograsp: Real-time EEG classification of high-level motor imagery tasks using a dual-stage deep learning framework," *IEEE Trans. Cybern.*, early access, Nov. 8, 2021, doi: 10.1109/TCYB.2021.3122969.
[53] C. Herff et al., "Brain-to-text: Decoding spoken phrases from phone representations in the brain," *Front. Neurosci.*, vol. 9, p. 217, Jun. 2015.
[54] M. Angrick et al., "Speech synthesis from ECoG using densely connected 3D convolutional neural networks," *J. Neural Eng.*, vol. 16, no. 3, 2019, Art. no. 36019.
[55] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," 2020, *arXiv:2004.05439*.

**Ji-Hoon Jeong** (Associate Member, IEEE) received the Ph.D. degree in brain and cognitive engineering from Korea University, Seoul, South Korea, in 2021.

He is currently an Assistant Professor with the School of Computer Science, Chungbuk National University, Cheongju, South Korea. His research interests include machine learning, brain–machine interface, and artificial intelligence.

**Jeong-Hyun Cho** received the M.S. degree in information and communication engineering from Ajou University, Suwon, South Korea, in 2017 with a graduate education agreement for military officers in South Korea. He is currently pursuing the Ph.D. degree with the Department of Brain and Cognitive Engineering, Korea University, Seoul, South Korea.

His research interests include haptic intelligence, machine learning, deep learning, and brain–computer interfaces.

**Byeong-Hoo Lee** is currently pursuing the Ph.D. degree with the Department of Brain and Cognitive Engineering, Korea University, Seoul, South Korea.

His research interests include deep learning, brain–computer interfaces, and signal processing.

**Seong-Whan Lee** (Fellow, IEEE) received the M.S. and Ph.D. degrees in computer science from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 1986 and 1989, respectively.

He is currently the Head of the Department of Artificial Intelligence, Korea University, Seoul, South Korea. His research interests include artificial intelligence, pattern recognition, and brain engineering.

Dr. Lee is a Fellow of the International Association of Pattern Recognition and the Korea Academy of Science and Technology.