**PAPER**

# Speech-imagery-based brain–computer interface system using ear-EEG

Journal of Neural Engineering

RECEIVED
18 August 2020

REVISED
24 November 2020

ACCEPTED FOR PUBLICATION
7 December 2020

PUBLISHED
23 February 2021

PAPER

# Speech-imagery-based brain–computer interface system using ear-EEG

**Netiwit Kaongoen**[1,2], **Jaehoon Choi**[1,2] and **Sungho Jo**[1,3]

[1] School of Computing, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
[2] Both authors contributed equally to this work.
[3] Author to whom any correspondence should be addressed.

**E-mail:** ghiejo10jo@kaist.ac.kr, basedseal@kaist.ac.kr and shjo@kaist.ac.kr

**Keywords:** brain–computer interface, ear-EEG, speech imagery, multilayer extreme learning machine

Supplementary material for this article is available online

## Abstract

*Objective.* This study investigates the efficacy of electroencephalography (EEG) centered around the user's ears (ear-EEG) for a speech-imagery-based brain–computer interface (BCI) system. *Approach*. A wearable ear-EEG acquisition tool was developed and its performance was directly compared to that of a conventional 32-channel scalp-EEG setup in a multi-class speech imagery classification task. Riemannian tangent space projections of EEG covariance matrices were used as input features to a multi-layer extreme learning machine classifier. Ten subjects participated in an experiment consisting of six sessions spanning three days. The experiment involves imagining four speech commands ('Left,' 'Right,' 'Forward,' and 'Go back') and staying in a rest condition. *Main results.* The classification accuracy of our system is significantly above the chance level (20%). The classification result averaged across all ten subjects is 38.2% and 43.1% with a maximum (max) of 43.8% and 55.0% for ear-EEG and scalp-EEG, respectively. According to an analysis of variance, seven out of ten subjects show no significant difference between the performance of ear-EEG and scalp-EEG. *Significance.* To our knowledge, this is the first study that investigates the performance of ear-EEG in a speech-imagery-based BCI. The results indicate that ear-EEG has great potential as an alternative to the scalp-EEG acquisition method for speech-imagery monitoring. We believe that the merits and feasibility of both speech imagery and ear-EEG acquisition in the proposed system will accelerate the development of the BCI system for daily-life use.

## 1. Introduction

Brain–computer interface (BCI) systems have been widely researched as an alternative method of communication and control for patients who have lost the ability to talk or move, such as those who suffer from locked-in syndrome (LIS) or amyotrophic lateral sclerosis (ALS) [1]. BCI systems work by translating the user's brain activities into computer or machine commands [1]. Many studies have proven that BCIs can successfully help those patients to regain their ability to live their normal life [2]. However, BCIs are not yet suitable for daily-life use. Apart from its efficacy, a BCI for daily life should be convenient, easy, fashionable, and harmonized with daily-life activities. BCI paradigms reported in many studies are limited by the mode of BCI. Reactive BCIs such as P300 [3] and steady-state visually evoked potential (SSVEP) [4] require stimuli from an external device (e.g. a monitor). This affects the wearability of a BCI and the visual stimuli might also induce user fatigue from staring at a monitor, making the reactive BCIs not optimal for daily-life use. While BCIs that use motor imagery (MI) [5] do not require a stimulus, it is limited by the degree of freedom when used for control of a computer or a machine. MI-based BCIs can also be unintuitive to use depending on the circumstances as users might find it difficult to relate MI tasks (e.g. imagining left-hand movement) to the task they want to be done (e.g. turning on the television).

To overcome these limitations, speech imagery has been researched and proposed as an alternative mode of BCI. Speech imagery is a type of mental task that refers to when a person imagines speaking aloud

without actually moving any articulators or speaking [6]. Speech-imagery-based BCI can be more intuitive compared to other types of BCI in that users can simply think of the word associated with the output command for the system to detect. Speech-imagery tasks also require less training time since most people are already naturally accustomed to it. It also, in theory, supports as many commands as there are sounds and combinations of sounds.

The majority of early speech-imagery research was based on imagining vowels. Fujimaki *et al* [6] first proposed the idea of speech imagery by examining the evoked potential from imagining the vowel /a/. In [7], DaSalla performed a classification of electroencephalography (EEG) signals acquired when subjects imagined the vowels: /a/ and /u/, or stay at rest using common spatial patterns (CSPs) from the collected EEG signals. Similarly, Matsumoto *et al* [8] used a support vector machine (SVM) and a relevance vector machine (RVM) to classify the EEG from imagining Japanese vowels. The study in [9] used Hilbert spectrum analysis to classify syllables: /ba/ and /ku/ imagined in different rhythms.

Speech imagery has also been researched using brain signals from electrocorticography (ECoG). Leuthardt *et al* [10] decoded phonemes using ECoG to control a one-dimensional cursor. The study in [11] classified both vowels and consonants with a Naïve Bayes classifier using ECoG signals for both overt and covert speech. They achieved an average accuracy of approximately 40% for both vowels and consonants in both overt and covert speech tasks. Martin *et al* [12] recorded ECoG responses for six words in three conditions: listening, imagined speech, and overt speech. They carried out pairwise classifications on 15 word pairs using an SVM. Of the 15 pairs, 8 showed accuracy significantly higher than the chance level for imagined speech.

Recently, speech imagery using words with semantic meaning has been researched as well. Nguyen *et al* [13] used Riemannian manifold features for the classification of short words, long words, and vowels imagined periodically at a fixed rhythm. Qureshi *et al* [14] performed classification using five imagined words ('Go,' 'Back,' 'Left,' 'Right,' and 'Stop') and achieved accuracy of up to 40.30%. García-Salinas [15] conducted both speech and visual imagery experiments for 13 words and images. They reported an accuracy of 34.2% and 26.7% for the speech and visual imagery experiments, respectively.

Most of these studies, however, were carried out with the conventional EEG acquisition methods that acquire EEG from the user's scalp using electrodes with electrolyte gel or electrical conductive paste (i.e. wet electrodes) on a cap. This method provides a high-quality EEG signal with a wide range of EEG channels that covers all parts of the human brain, which makes this EEG acquisition technique often the best non-invasive technique when it comes to the accuracy of BCI systems. Nevertheless, scalp-based EEG acquisition methods are unsuitable for BCIs that are intended to be used in everyday life for three main reasons: (a) the equipment preparation takes time and requires extra training to learn the procedure, (b) the cap and wet electrodes make them uncomfortable, and (c) they are not fashionable and can be socially awkward. To solve these problems, researchers have tried replacing the wet electrodes with different types of electrodes and changing the design of the EEG acquisition tool to make it wearable. Examples of commercial-grade wearable EEG acquisition devices include NeuroSky (www.neurosky.com), which uses one active dry electrode to acquire EEG from the user's forehead, and Emotiv (www.emotiv.com), which uses disposable sponge electrodes with saline solution.

EEG centered around the user's ears (ear-EEG) is an alternative EEG acquisition method that has been gaining popularity in the field of BCI research due to its comfortability, mobility, and discreetness. This EEG acquisition method measures EEG centered around the user's ears. Ear-EEG does not require any complicated equipment preparation and the sensors do not have any contact with the user's hair; thus, it is easier and more comfortable for users to use than the conventional scalp-EEG methods. The electrode placement in ear-EEG methods also makes them invisible to other people and does not attract any unwanted attention to the user, making ear-EEG methods very discreet and suitable for daily life. Looney's research [16] is one of the first to propose the concept of ear-EEG. They developed orthoplastic earpieces that acquire EEG from the inside of users' ear canals. Debener's group [17] took a different approach that acquired ear-EEG from around the ear. They developed an around-ear EEG acquisition tool called cEEGrid that consists of ten electrodes printed on a C-shape flexible sheet. Following these works, many studies have developed their own ear-EEG acquisition tools and shown that ear-EEG is a reliable data acquisition method for BCI systems. The BCI signal types that can be detected by ear-EEG include alpha attenuation [18], auditory steady-state response [18], concentration level [19], auditory attention state [20, 21], sleep state assessment [22], SSVEP [23], and auditory event-related potential [18, 24, 25].

In order to accelerate the development of daily-life BCIs, we propose a speech-imagery-based BCI system using ear-EEG. In this study, we develop a wearable and low-cost around-ear-EEG acquisition device and investigate the efficacy of the ear-EEG acquisition method in speech-imagery-based BCI systems. We measure EEG from the scalp and ears simultaneously in a multi-class speech-imagery experiment and directly compare the classification results between the two EEG acquisition methods. Furthermore, a model is also trained to map the

ear-EEG features into the scalp-EEG feature space in an attempt to improve the accuracy of the ear-EEG-based system. Our feature extraction method is based on the EEG covariance matrices in the Riemannian framework. We use a multi-layer extreme learning machine (MLELM) classifier as the classification method in our system. The methods used in our system are described in detail in the following section. To the best of our knowledge, this study is the first to use ear-EEG as the data acquisition method for a speech-imagery-based BCI.

## 2. Method

### 2.1. Data acquisition

*2.1.1. Ear-EEG*

We measure ear-EEG signals from around both of the subject's ears in six channels, three from each side. The channel names are L1, L2, and L3 for the electrodes around the left ear and R1, R2, and R3 for the ones around the right ear. The signals are referenced and grounded to the electrodes at the bottom of the right (REF) and left ear (GND), respectively. The electrodes on each side are arranged in a C-shape 55 mm in height and 20 mm in width. Figure 1(a) illustrates the position of the ear-EEG channels.

Ear-EEG is acquired using low-cost wearable equipment custom-made for this study. The wearable equipment is in the shape of a horizontal headband that covers the back of the user's ears and wraps around the back of the user's head. The equipment contains the C-shape earpieces made with flexible silicone (Dragon Skin 30) that cover around the ears. We use foam-type solid-gel snap electrodes (3 M Red Dot) cut to 14 mm in diameter for our equipment. The electrodes can be easily attached and detached from the sockets embedded in the silicone earpieces. The silicone earpieces and foam-type snap electrodes give a soft touch to the user's skin, which makes the device comfortably wearable. The electrodes achieve impedance below 15 kΩ, which is similar to that of the cEEGrid [17] without applying any extra electrical conductive substance. The electrodes do not dry out and remain at the same impedance level for at least 6 h.

The silicone earpieces are attached to a 3D-printed frame made with ABS material. The wires are connected to an EEG sensing board contained in a 3D-printed case that can be hooked to the user's clothes. The case also contains a portable battery and a charger. We use OpenBCI's Cyton Biosensing Board (www.OpenBCI.com) as the EEG sensing board. The EEG-acquisition sampling rate is 250 Hz. The battery lasts for at least 10 h for continuous EEG recording. Figure 2 shows pictures of our ear-EEG wearable device.

The low-cost wearable ear-EEG equipment is discreet and can be worn comfortably. The equipment is well concealed compared to the EEG cap and other scalp-based wearable tools. The equipment can be constructed easily by hand with the help of a 3D printer that makes the frame, and all materials are available commercially. The design of our equipment also allows for a wide range of applications. For example, the headband frame could be modified to attach a camera or other sensors that can be used to target an object in the environment for control. The equipment setup process includes peeling the sticker out of the electrodes and attaching them to the sockets in the silicone parts of the equipment. For paralysis patients, this process can be done easily with help from an extra person without any special training. The total equipment preparation time of our ear-EEG acquisition tool is less than 3 min.

*2.1.2. Scalp-EEG*

We acquire the scalp-EEG at a sampling rate of 500 Hz using BrainVision actiCHamp with an EEG cap consisting of 32 Ag/AgCl electrodes placed around the left hemisphere following the 10–20 international system (figure 1(b)). Fpz and FCz are chosen as the ground and reference channels, respectively. Broca's area (F5, FT7, FC5, and FC3) and Wernicke's area (TP7, CP5, CP3, and P5) are associated with language production and comprehension, respectively [26, 27], and it has been demonstrated in previous studies that the brain activities in these areas are dominant during speech-imagery tasks [13, 26, 28]. Thus, instead of spanning the electrodes across the user's scalp, we place the electrodes densely on the left hemisphere so that the chance of picking up meaningful data during the speech imagery is maximized while maintaining the number of electrodes as 32. This could shorten the equipment preparation time by half when compared to the 64-channel setup that densely covers all areas of the user's scalp. Electrodes are not placed on channels T9, TP9, and P9 due to their proximity to the ear-EEG device. Electrolyte gel is inserted to ensure the connection between the electrodes and scalp and keep the impedance level below 10 kΩ. The scalp-EEG equipment preparation takes approximately 30 min to complete.

### 2.2. Experimental setup

All experiments were carried out in a soundproofed room to minimize external sound noise. Each subject performed the experiment for a total of six sessions spanning over three different days, two sessions a day, with 20 min rest time between two sessions, and each session of the experiment taking approximately 20 min. Subjects were prepared for EEG acquisition in a comfortable chair about a meter away from a large monitor.

Experimental procedures were explained at the start of the experiment with visual cues. We gave the subjects ample time to practice and encouraged them to ask questions to ensure that they completely understood the tasks. We recorded both ear-EEG and

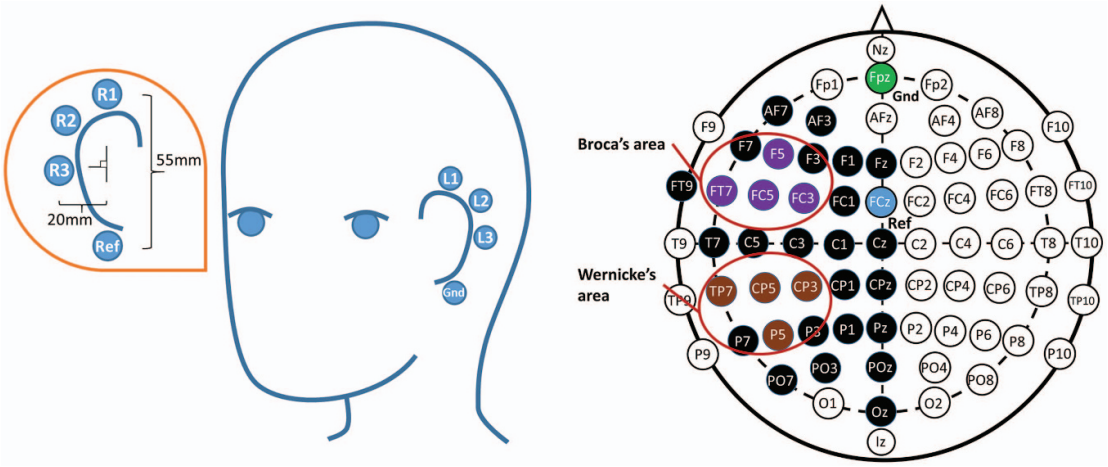


**Figure 1.** Electrode placement of (a) ear-EEG and (b) scalp-EEG used in this study.


**Figure 2.** Low-cost wearable ear-EEG equipment (a). Subject wearing the equipment in side-view (b). Ear-EEG electrodes are wired to the case containing the sensing board, which can be hooked to the user's clothes (c).

scalp-EEG simultaneously during the experiments. Due to the EEG cap, we removed the 3D-printed frame from the ear-EEG device and used the silicone earpieces only. There were ten blocks of tasks in each session of the experiment. Each block contained four speech-imagery tasks for the speech commands 'Right,' 'Left,' 'Forward,' and 'Go back,' and a control task in which the subjects were asked to relax with eyes open (labeled as 'Rest'), in a randomized order. Each task contained five trials of the speech imagery or resting state, comprising a total of 50 trials for each task.

A speech-imagery task started with an audio cue, where the corresponding word was read by a female voice in an American accent. After 2 s, a crosshair cue was shown for 1 s during which subjects were instructed to relax. Then, a circle sign was given for 2 s, during which they were expected to pronounce the speech command given before. Following this, 1 s of crosshair was shown again for relaxation. Actual articulated speech generally took less than a second, so the subjects actually had more than 1 s to rest for the next step. A loading bar was then shown for 2 s, during which subjects were instructed to imagine the speech command in a stretched-out manner according to the progress of the loading bar. This was shown five times in a row, with 1 s of crosshair shown in between. Subjects were given 2.5 s to rest afterward before the next task started. The control tasks were carried out in a similar manner to the speech-imagery tasks but a beep sound was given instead of the audible word and there was no following step for the articulated speech. In this task, subjects were instructed to look at the loading bar without imagining any speech. The experimental procedures are illustrated in figure 3.

### 2.3. Data pre-processing

We first apply a notch filter with 60 Hz cutoff frequency to the raw scalp-EEG and ear-EEG data from each session of the experiment to remove the noise from the power line. The EEG data are then segmented into multiple 2 s EEG epochs for each trial starting from the onset of the visual cue and labeled with their corresponding class. Finally, we decompose the EEG epochs into five different EEG frequency bands including delta (0.5–4 Hz), theta (4–7 Hz), alpha (7–14 Hz), beta (14–30 Hz), gamma (30–100 Hz) and 'Board' (0.5–100 Hz) using a fourth-order Butterworth band-pass filter. The first five frequency bands are common EEG frequency bands that are categorized according to their unique characteristics and functions, and the purpose of the 'Broad' band is to capture and process the EEG data as a whole. By decomposing the EEG data into different bands and extracting the features, we can analyze and use the results of the experiment to investigate the relationship between the EEG and cognitive mechanisms of speech-imagery tasks.

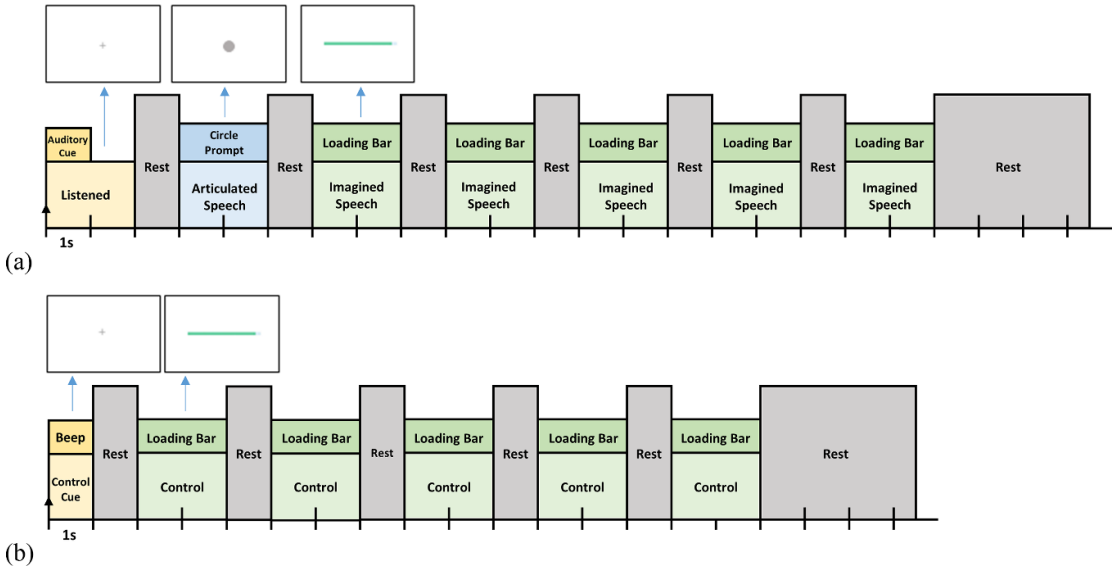


**Figure 3.** Experimental procedure for (a) speech-imagery tasks (speech imagery of the commands: 'Right,' 'Left,' 'Forward,' and 'Go back') and (b) control (resting-state) task. Each task contains five trials of the speech imagery or resting state. This procedure is repeated ten times for each task in a box-randomized manner, comprising a total of 50 trials per each task.

### 2.4. Feature extraction

*2.4.1. Covariance matrix*

From the data pre-processing step, the EEG data epochs of each frequency band of the $i$th trial can be represented as a matrix $X_i = [x_1, \ldots, x_T] \in R^{n \times T}$ where $n$ denotes the number of channels and $T$ is the number of data points in an epoch. The covariance matrix $P_i \in R^{n \times n}$ is defined as:

$$P_i = \frac{1}{T-1} X_i X_i^T. \tag{1}$$

The result covariance matrix $P_i$ is a symmetric positive-definite (SPD) matrix.

*2.4.2. Tangent space projection of a covariance matrix*

Since the spaces of SPD matrices lie in the Riemannian manifold, we cannot effectively use covariance matrices directly as the features for the classification algorithms that are based on projections into hyperplanes [29]. In this study, we project the covariance matrices into their corresponding tangent space and construct the tangent vectors to make them effectively usable as the features for the classification algorithms. For each covariance matrix $P_i$, its tangent space vector ($s_i \epsilon R^m$, where $m = \frac{n(n+1)}{2}$) is defined as:

$$s_i = \text{upper}\left(P_R^{-\frac{1}{2}} \log_{P_R}(P_i) P_R^{-\frac{1}{2}}\right) \tag{2}$$

where upper $(X)$ is the operator to keep only the upper triangular part of the matrix $X$ and vectorize it by applying the unity weight to the diagonal elements and $\sqrt{2}$ weight to the others, $P_R$ is the Riemannian mean of the $N$ covariance matrices, and $\text{Log}_C(P)$ is the logarithmic mapping of matrix $P$ using the reference point $C$ defined as:

$$\text{Log}_C(P) = C^{\frac{1}{2}} \log\left(C^{\frac{1}{2}} P C^{\frac{1}{2}}\right) C^{\frac{1}{2}}. \tag{3}$$

The detailed descriptions of the Riemann geometry properties of the SPD matrices and the tangent space projection process can be found in [29].

In our system, the Riemannian tangent space vectors of covariance matrices are calculated separately for each frequency band. The feature matrix is then constructed by concatenating all tangent vectors from each frequency band. For the sake of simplicity, we label this feature extraction method as TS. The dimension of the feature matrix is $(N_s \times 126)$ for ear-EEG and $(N_s \times 2976)$ for scalp-EEG, where $N_s$ is the number of samples. Finally, analysis of variance (ANOVA) $F$-values are calculated and used to select the best $k$ features for classification, making the dimensions of the final feature matrix $(N_s \times k)$. We run the algorithm using $k = [1, 10, 20, \ldots, 110]$ for ear-EEG and $k = [1, 100, 200, \ldots, 2500]$ for scalp-EEG. The feature selection method can help reduce the computational cost for the system and might improve the accuracy of the system. In section 3, we show the effect of the feature selection method on the classification result and discuss the features that are most significant according to the $F$-test ANOVA.

### 2.5. Classification method

*2.5.1. Extreme learning machine*

An extreme learning machine (ELM) is a single-layer feed-forward neural network that consists of an input layer, a single hidden layer, and an output layer [30]. The difference between ELM and common neural networks is that the hidden layer does not need to be

tuned. The weights of the input layer and the bias values of the hidden node are assigned randomly and are not learned or updated throughout the process. ELM is extremely fast in training due to its random initialization for the input weights and bias values, which makes it suitable for BCI in daily-life applications where the classification model should be updated regularly because of the non-stationary nature of EEG. ELM also shows better performance in speech-imagery-based BCI compared to other common classification methods in previous studies [13, 14].

For $N$ distinct samples $(x_i, y_i)$, $x_i \in R^{N\times j}$ and $y_i \in R^{N\times m}$ where $i = 1,\ldots,N$, $j$ is the number of input nodes, and $m$ is the number of output nodes, the hidden layer output is defined as:

$$h(x_i) = g(ax_i + b) \tag{4}$$

where $g(x)$ is the activation function, $a \in R^{j\times n}$ is the input weights and $b \in R^n$ is the bias. The output layer can then be expressed as:

$$h(x_i)\, V = y_i \tag{5}$$

where $V \in R^{n\times m}$ is the matrix of output weights. Considering all $N$ training samples, the ELM model with $n$ hidden nodes can be constructed as:

$$HV = Y \tag{6}$$

where

$$H = \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_n) \end{bmatrix} = \begin{bmatrix} h_1(x_1)\cdots h_n(x_1) \\ \vdots\ddots\vdots \\ h_1(x_n)\cdots h_n(x_n) \end{bmatrix}_{N\times n},$$

$$V = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}_{n\times m}, \text{ and } Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{N\times N}.$$

There are three basic steps to learn the ELM model. The first step is to assign a random value (between 0 and 1) to the input weights $a$ and bias $b$. The second step is to calculate the matrix $H$. Finally, the output weight can be calculated as $V = H^{\mathrm{t}}Y$ where $H^{\mathrm{t}}$ is the Moore–Penrose generalized inverse of matrix $H$.

Auto-encoding ELM (ELM-AE) is one variation of the ELM model. ELM-AE is an unsupervised learning ELM that is constructed by having the output of the ELM network the same as the input of the networks and trained in the same manner as a normal ELM model.

*2.5.2. MLELM*

MLELM is a deep-learning variation of an ELM. It is constructed by using multiple ELM-AEs [31] to train the input for each hidden layer. Figure 4 depicts the structure of an MLELM model with $k$ hidden layers. As seen from the figure, the $l+1$th hidden layer is constructed by an ELM-AE that takes the $l$th hidden layer ($h_l$) as the input (figure 4(a)). The output weight $V_l$ learned from the ELM-AE is then used to transfer the $l$th hidden layer to the higher level of feature space (figure 4(b)). Mathematically, the $l$th hidden layer of an MLELM model can be expressed as:

$$H_l = g\left((V_l)^{\mathrm{T}} H_{l-1}\right). \tag{7}$$

It should be noted that in the first hidden layer ($l = 1$), $H_0$ is the input layer $x$. The output weights that connect the last hidden layer and the output layer are then learned in the same way as the original ELM.

### 2.6. Ear-to-scalp feature mapping

We hypothesize that the scalp-EEG will give a better result than the ear-EEG, so we attempt to improve the result from ear-EEG by mapping the ear-EEG feature matrix to the scalp-EEG feature space (labeled as the EtoS method).

The mapping process is done using an ELM model. The EtoS model is trained in the same way as in the classification task but, instead of setting the sample label $y$ as the output layer, we use the scalp-EEG features of the same sample as the output layer. In this study, the number of hidden nodes is set to 2976, the same number as the scalp-EEG features. The EtoS feature matrix is further processed and classified in the same way as the ear-EEG and scalp-EEG feature matrix. Figure 5 summarizes the procedure of our proposed BCI system.

### 2.7. Methods from previous studies

Other than using the TS feature extraction method with the MLELM classifier (labeled as TS + MLELM), we also process and classify the EEG data from speech-imagery tasks using methods presented in previous BCI studies to compare and evaluate the performance of our methodology. Several classifiers including linear discriminant analysis, linear SVM, ELM, and RVM are used as the classifier with the TS feature extraction method. RVM is similar to SVM but uses a Bayesian framework to obtain the sparse solutions [32]. The study in [13] shows that RVM is superior to other classifiers including ELM in classifying speech-imagery tasks. We also perform different approaches that have been proven to be effective in MI-based BCI systems, which include using the filter bank CSP (FBCSP) as the feature extraction method with an SVM classifier (labeled as FBCSP + SVM) [33], and using pre-processed EEG data directly as an input feature to ShallowNet (labeled as EEG + ShallowNet) [34]. In the FBCSP + SVM method, we band-pass filter the EEG signal into five main frequency bands (delta, theta, alpha, beta, and gamma) and extract six CSP features from each frequency band. ShallowNet is a convolution neural network with a shallow structure,

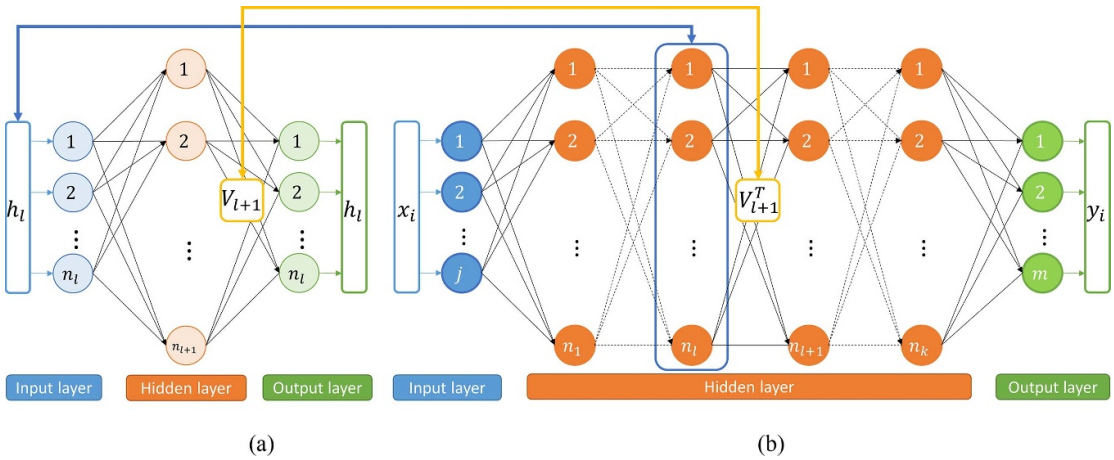


**Figure 4.** Model structure of (a) an ELM-AE and (b) an MLELM.

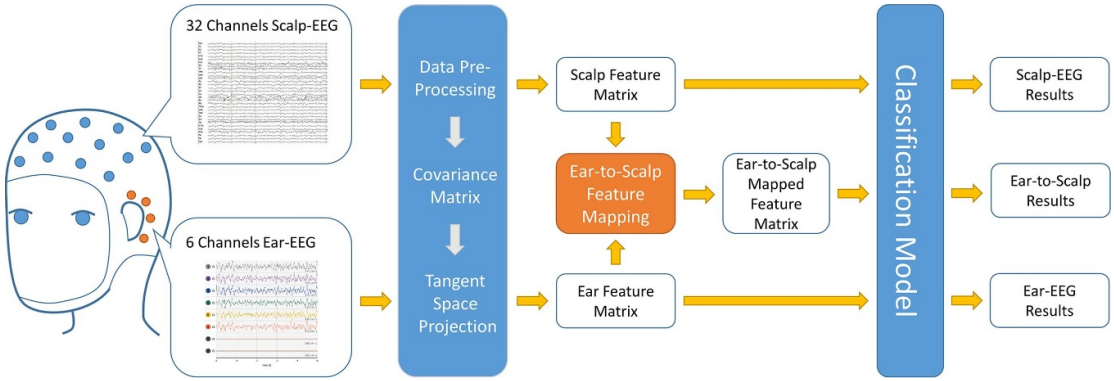


**Figure 5.** Procedure of the proposed system for EEG data acquisition, feature extraction and classification.

which has previously been reported to increase accuracy compared to the FBCSP + SVM method in MI-based BCIs [34]. We construct our ShallowNet in the same way as described in [34].

In addition, we use the upper triangle of covariance matrices directly as input features to an MLELM classifier. This method is labeled as COV + MLELM.

### 2.8. Participants

In this study, ten male subjects, 20–29 years of age and fluent in English, were recruited. All of the subjects were free from any neurological disorders and reported no visual and hearing impairments or significant health problems. Four subjects had no prior experience in BCI while the other six subjects had previous experience participating in BCI experiments but not in one that is based on speech imagery. All subjects gave written informed consent. The KAIST Institutional Review Board approved the proposed experimental protocol of this study.

### 2.9. EEG visualization

To better understand the characteristic of EEG activities during speech-imagery tasks, we visualize the EEG data obtained during all tasks in both spectral and time-frequency domains. Spectral analysis is carried out by acquiring the power spectrum density (PSD) of each 2 s EEG epoch using the multitaper method. Then, we perform an $F$-test on the acquired PSD values from each speech-imagery task in comparison to the resting state to obtain the corresponding $F$-values, which helps us gain more knowledge of what spectral and spatial features are dominant during speech-imagery tasks.

Time-frequency analysis is conducted using the Morlet wavelet transform. In this analysis, we divide the EEG channels into different groups to examine the characteristics of EEG during speech-imagery tasks in each specific area of interest. Ear-EEG channels are divided into two groups: left-ear (L1, L2, and L3) and right-ear (R1, R2, and R3), and scalp-EEG channels are divided into four groups: Broca's area (F5, FT7, FC5, and FC3), Wernicke's area (TP7, CP5, CP3, and P5), midline sagittal plane (Fz, Cz, CPz, Pz, POz, and Oz) and temporal channels (T7 and FT9). Broca's and Wernicke's areas are chosen for their association with speech functions and the temporal channels are chosen for their proximity to

the ear-EEG channels. Time-frequency responses are averaged across the channels in each group for three conditions: speech imagery of short speech commands ('Left' and 'Right'), speech imagery of long speech commands ('Forward' and 'Go back'), and rest condition.

### 2.10. System evaluation

The system is evaluated using a ten-fold cross-validation for each session of the experiment. This gives 225 training samples and 25 testing samples for each iteration of cross-validation. The cross-validation is performed in such a way that the samples from the same block stay in the same fold. The tangent space projector and feature selector are computed using only the data from the training samples. We use the grid search method to find the optimized number of hidden nodes from $[50, 60, \ldots, 200]$ in each layer of the ELM and MLELM model for the cross-validation. Since the ELM model and its variations use randomized values as their weight and bias values, the random seed is specified so that each run of the model training gives the same output. The accuracy results from all ten iterations of the cross-validation are averaged to represent the accuracy result of each session of the experiment. Moreover, confusion matrices are calculated for each session to examine the prediction accuracy of each class as well.

In addition to the comparisons between results from the ear-EEG, scalp-EEG, and EtoS methods, and a comparison between our method and methods from previous studies, we also examine our system in other aspects. First, we compute and compare the classification results obtained from the ear-EEG data from all subjects in three channel-settings: using both left and right channels, using only the left channels, and using only the right channels. This helps us gain more knowledge of the performance of the ear-EEG in different channel settings in the speech-imagery-based BCI. Finally, we investigate the effect of training on a user's performance in the speech-imagery task by comparing the classification results between each session of the experiment to see if there is any improvement in the performance as the subjects gain more experience with the experiment. All result comparisons are carried out using $t$-tests to find out their statistical significance. In multiple comparisons (i.e. comparison between the results of ear-EEG and scalp-EEG from each subject), the Bonferroni method is used to correct the confidence level of the $p$-value. The performance of the proposed system and the data analysis are shown and discussed in the next sections.

## 3. Results

### 3.1. Data analysis

#### *3.1.1. Data visualization*

Our data visualizations show that each participant has different patterns of brain activities during the speech-imagery tasks, yet there are some underlying similarities between some subjects. Figures 6 and 7 show the spectral analysis and time-frequency analysis of both ear-EEG and scalp-EEG for subject S04, respectively. The EEG data of subject S04 are chosen to be presented here due to their high classification results in both ear-EEG and scalp-EEG (shown in section 3.2). Data visualizations for other subjects are provided in the supplementary data (available online at stacks.iop.org/JNE/18/016023/mmedia). It should be noted that the following observations described in this section are specific to the data from subject S04 and may not be applicable to the data from other participants.

In the spectral analysis of ear-EEG (figure 6(a)), we see relevant activity from 20 Hz onward. Of the four classes, the 'Right' class shows the lowest $F$-values. This is also shown in the data from subjects S02, S03, S04, S08, and S10 (supplementary data A). The 'Left' and 'Forward' classes show higher $F$-values especially from the left ear when compared to the other classes. The 'Go back' class displays high $F$-values only in the R1 channel. The R3 channel shows little to no difference across all classes, possibly due to its proximity to the reference electrode. This can be seen in most of the subjects except for S01, S07, and S10 (supplementary data A). For scalp-EEG, different trends can be seen for each speech command in the spectral analysis (figure 6(b)). The 'Left' class shows high $F$-scores in the channels around Broca's area, Wernicke's area, and the temporal channels (T7 and FT9) from 30 Hz onwards. This applies to subjects S03, S06, S07, and S10 as well (supplementary data B). The 'Right' class shows more similarity to the rest condition in PSD when compared to the 'Left' class, except in subjects S01, S02, S05, and S09. The 'Forward' and 'Go back' classes show similar $F$-scores to each other. These long speech commands show activity focused in Broca's area, but less so in Wernicke's area.

In the time-frequency analysis (figure 7(a)), we can see that the ear-EEG from the left ear shows higher activations compared to the right ear in speech-imagery tasks. While the response for short speech appears from 0.25 s onset, long speech shows lower and more delayed responses. Delayed responses can also be observed from the data of subjects S01, S03, S08, S09, and S10 (supplementary data C). Activities above 30 Hz can be seen from both short and long speech imagery in most of the subjects except for S01, S02, and S06. In figure 7(b), Wernicke's area shows the highest activity during both short and long speech imagery compared to other areas at above 30 Hz: 0.3 s from the onset of the loading bar for short commands, and 0.5 s for long commands. The delayed activities for long commands are also shown in subjects S03, S07, S08, and S10 (supplementary data D). Broca's area and the temporal channels also show a similar pattern to Wernicke's area for short words, but

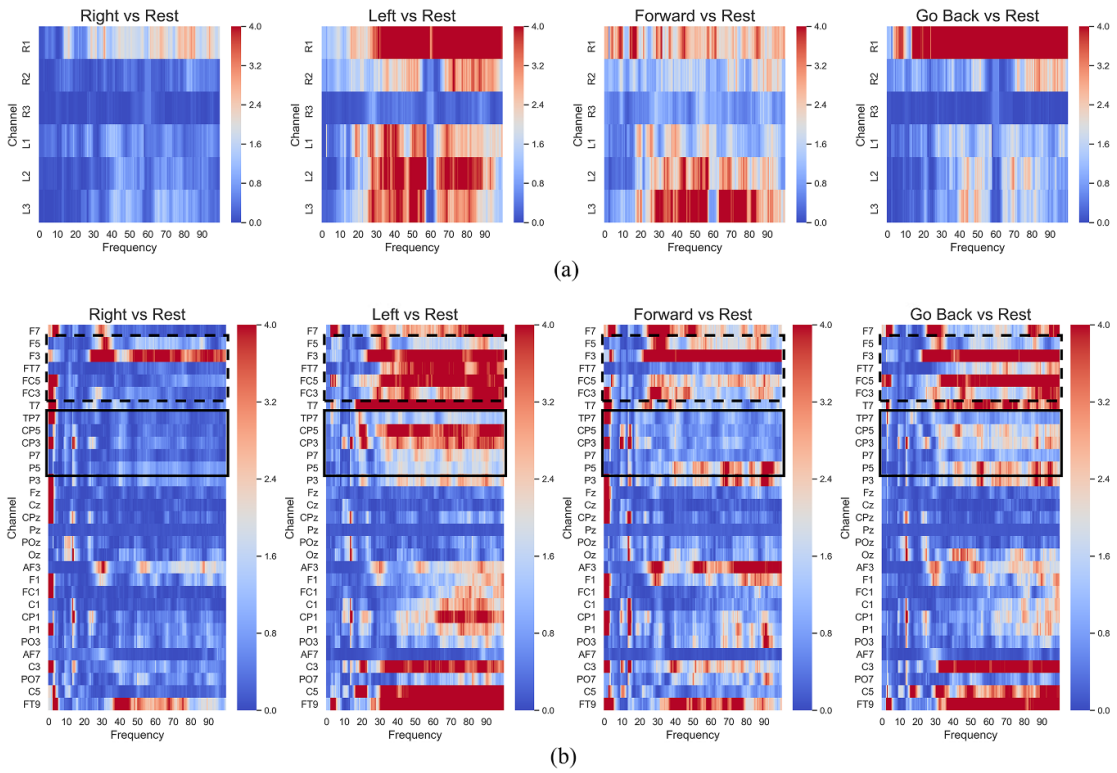


**Figure 6.** *F*-scores of power spectrum density calculated from each speech-imagery class vs 'Rest' class of (a) ear-EEG and (b) scalp-EEG for subject S04. The dotted box indicates the channels in Broca's area and the solid box indicates the channels in Wernicke's area. The 'Left' class shows the highest *F*-value in both methods. All speech-imagery tasks are most dominant in Broca's area. A higher value of *F*-score is shown in hot color (red) as indicated on the color scale.

at a lower amplitude. Broca's area also shows activity in the frequency below 10 Hz and the temporal channels display some activities during the control task.

*3.1.2. Feature analysis*

We further investigate the characteristics of EEG during the speech-imagery tasks by analyzing the input features. In this analysis, we perform the ANOVA *F*-test separately for each session of the experiment without dividing the data into training and testing sets as we have done for the actual feature selection process.

Figure 8 shows the averaged *F*-score (*y*-axis) of each feature (*x*-axis) across all sessions for ear-EEG data (a) and scalp-EEG (b). We also calculate the mean *F*-score for the features from each frequency band. From ear-EEG data, we can see that the features from the gamma band have the highest mean scores followed by the 'Broad' and delta band, while the theta band has the lowest mean *F*-score. For scalp-EEG, gamma features again have the highest mean *F*-score and the theta band has the lowest *F*-score.

### 3.2. Classification result

*3.2.1. Comparison between ear-EEG and scalp-EEG results*

Table 1 shows the classification results of our system. The mean accuracies and standard deviations (std) are derived using the results of all six sessions for each subject. The average accuracy for ear-EEG and scalp-EEG when using all features is 37.3 ± 3.2% and 41.9 ± 6.4%, respectively. Table 1(b) shows classification results and the best *k* number for each subject when the feature selection method is applied. The average accuracy for ear-EEG and scalp-EEG across all subjects, in this case, is 38.2 ± 3.3% and 43.1 ± 6.5%, respectively. When the *k* number is fixed for all subjects, the result shows a very small improvement in the average accuracy across all subjects (best result: 37.6%, $k = 50$ for ear-EEG and 42.8%, $k = 1000$ for scalp-EEG). The feature selection method does not significantly improve the accuracy of the system ($p > 0.5$ for both ear-EEG and scalp-EEG).

The results show that the classification accuracies of all sessions are significantly higher than the chance level (20%) in both ear-EEG and scalp-EEG methods (one-tailed *t*-test, $p < 0.01$). The maximum (max) and minimum (min) results are 43.0% (subject S01) and 32.9% (subject S09) for ear-EEG, and 55.0% (subject S03) and 36.1% (subject S09) for scalp-EEG. When comparing the results of scalp-EEG and ear-EEG for each subject, only the scalp-EEG results of subjects S02, S03, and S07 are significantly better than the ear-EEG result ($p < 0.001$), while the other seven subjects show no significant increase in classification result from scalp-EEG. The difference between ear- and scalp-EEG results is higher than 10% only in subjects S02 and S03. Furthermore, subject S05

**Table 1.** Mean ± std of the accuracy (%) of the proposed system using ear-EEG and scalp-EEG for all subjects.

| | Subject | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| EEG-type | S01 | S02 | S03 | S04 | S05 | S06 | S07 | S08 | S09 | S10 | Average |
| (a) Using all features | | | | | | | | | | | |
| Ear | 41.9 ± 3.4 | 39.0 ± 2.4 | 41.5 ± 2.8 | 38.4 ± 2.7 | 36.6 ± 4.1 | 35.0 ± 6.1 | 40.0 ± 16 | 35.6 ± 3.2 | 31.9 ± 1.8 | 33.1 ± 2.0 | 37.3 ± 3.2 |
| Scalp | 44.6 ± 2.5 | 49.0 ± 4.0 | 53.8 ± 2.6 | 47.9 ± 6.6 | 36.1 ± 2.9 | 36.2 ± 5.5 | 43.3 ± 17 | 36.4 ± 2.8 | 35.3 ± 3.5 | 36.1 ± 3.2 | 41.9 ± 6.4 |
| EtoS | 41.7 ± 3.6 | 40.9 ± 1.7 | 43.6 ± 1.9 | 38.4 ± 3.8 | 35.9 ± 3.6 | 35.0 ± 4.4 | 40.6 ± 17 | 34.9 ± 3.5 | 32.5 ± 2.4 | 33.1 ± 2.6 | 37.7 ± 3.7 |
| (b) Feature selection (best $k$ number) | | | | | | | | | | | |
| Ear | 43.0 ± 3.9 (60) | 40.6 ± 2.7 (30) | 42.5 ± 2.6 (70) | 38.9 ± 3.6 (100) | 38.2 ± 3.3 (50) | 35.6 ± 5.6 (100) | 40.1 ± 1.6 (70) | 35.7 ± 4.1 (50) | 32.9 ± 3.1 (80) | 34.1 ± 2.2 (100) | 38.2 ± 3.3 |
| Scalp | 46.6 ± 1.1 (1000) | 51.3 ± 4.3 (1500) | 55.0 ± 3.2 (1200) | 48.3 ± 7.0 (1200) | 37.9 ± 3.8 (1000) | 36.8 ± 4.5 (1000) | 43.4 ± 0.9 (900) | 38.1 ± 2.5 (600) | 36.1 ± 4.3 (1300) | 37.7 ± 3.7 (1000) | 43.1 ± 6.5 |
| EtoS | 43.0 ± 4.3 (900) | 42.1 ± 2.1 (900) | 43.8 ± 1.8 (300) | 38.7 ± 3.7 (500) | 36.1 ± 2.9 (700) | 35.2 ± 4.2 (1200) | 41.6 ± 1.7 (500) | 35.4 ± 3.4 (1200) | 32.9 ± 3.0 (100) | 33.5 ± 2.9 (900) | 38.2 ± 3.9 |

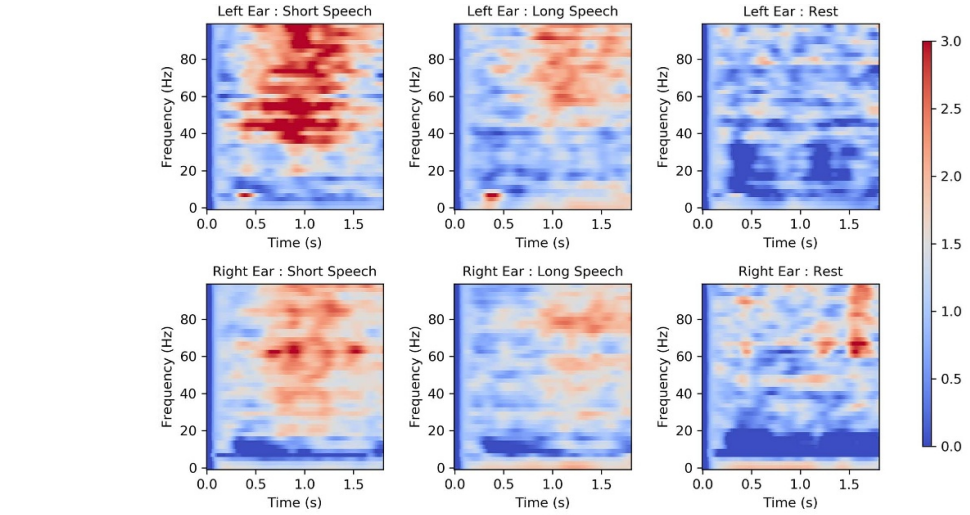


**Figure 7.** Time-frequency analysis of (a) ear-EEG and (b) scalp-EEG data in different channel groups for subject S04 averaged across all sessions. Time-frequency analysis was carried out using the Morlet wavelet transform. Channels were divided into two groups for ear-EEG: left-ear (L1, L2, and L3) and right-ear (R1, R2, and R3), and four groups for scalp-EEG: Broca's area (F5, FT7, FC5, and FC3), Wernicke's area (TP7, CP5, CP3, and P5), midline sagittal plane (Fz, Cz, CPz, Pz, POz, and Oz) and temporal channels (T7 and FT9). Greater power is shown in hot color (red) as indicated on the color scale.

even shows a slightly higher result from ear-EEG than scalp-EEG (ear-EEG result = 38.2 ± 3.3%, scalp-EEG result = 37.9 ± 3.8%, $p = 0.78$).

### *3.2.2. Mapping ear-EEG features into the scalp-EEG feature space*

The classification result of the EtoS method is also shown in table 1. At the beginning of the experiment, we hypothesized that scalp-EEG would produce a much better classification result than the ear-EEG; thus, mapping the ear-EEG feature to scalp-EEG feature space might improve the performance of the ear-EEG. However, we can see from the results that both ear-EEG and EtoS methods have the same classification accuracy averaged across all subjects when using the feature selection method, and the EtoS method has an average accuracy 0.4% higher without using the feature selection method. Of the ten subjects, only subjects S02, S03, and S07, whose scalp-EEG results are significantly better than their ear-EEG

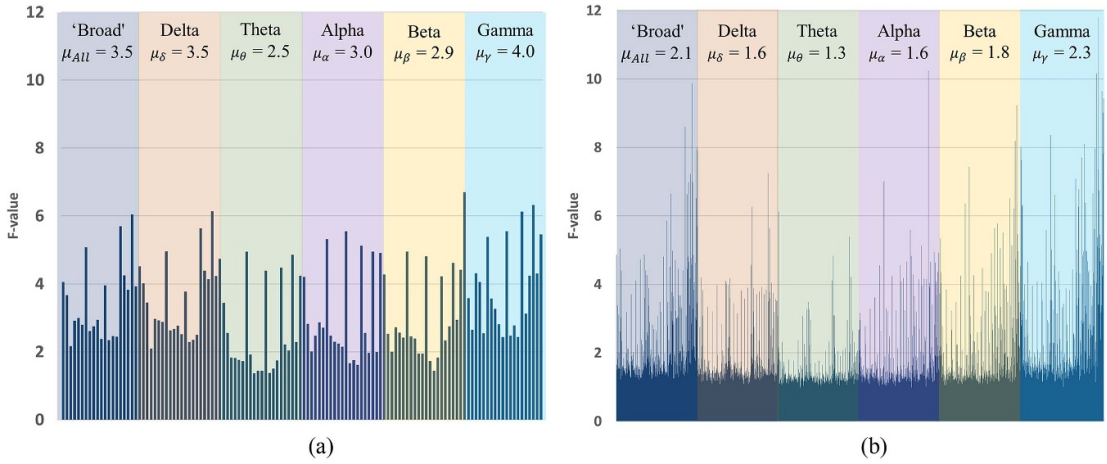


**Figure 8.** *F*-value of all (a) ear-EEG and (b) scalp-EEG features in each frequency band averaged across all sessions. Gamma band features show the highest significant level in both cases.

results, show a small increment in mean accuracy from the EtoS method. Nevertheless, the increments are not statistically significant in any of these subjects ($p > 0.05$).

*3.2.3. Ear-EEG: left vs right channels*

To further explore the setup of the ear-EEG acquisition methods, we compare the classification results obtained from the ear-EEG data from all subjects in three channel-settings. The classification process for this comparison is performed without using the feature selection method. The classification accuracies (mean ± std) are 37.3 ± 3.2%, 36.38 ± 3.6%, and 34.4 ± 2.9% for using both left and right channels, using only left channels, and using only right channels, respectively. Using *t*-tests, we found two interesting observations. First, the classification results from using only the right channels are significantly lower than when using only the left channels ($p < 0.05$) and when using both the left and right channels ($p < 0.01$). Second, the classification results from using all the left and right channels are not significantly different from when using only the left channels ($p > 0.05$).

*3.2.4. Confusion matrix*

From the data analysis, we observe that different kinds of speech commands have different patterns in brain activity in EEG acquired during the speech-imagery task. To further investigate this matter, we obtain the confusion matrix from the classification result of the TS + MLELM method averaged across all sessions (figure 9). The 'Rest' class shows the highest true positive rate in both the ear-EEG (52.0%) and scalp-EEG (62.6%) methods. Among the four speech-imagery classes, the 'Left' class has the highest true positive rate in both EEG types (36.5% and 42.1% for ear-EEG and scalp-EEG, respectively). The 'Right' class trials are misclassified as the 'Rest' class most frequently. Another interesting observation from both confusion matrices is that the samples from the long-speech commands are most frequently misclassified as each other. We can also see that the same applies to the short-speech classes for the ear-EEG, but the 'Right' class is misclassified most frequently as the 'Go back' class in scalp-EEG samples.

### 3.3. Comparison with methods from previous studies

To further evaluate the effectiveness of our system, we compare the classification results of our method (TS + MLELM) with several methods used in previous BCI studies (table 2). The mean, std, max, and min values are taken from the classification results of all sessions from all subjects. The results show that the MLELM classifier significantly outperforms all other classifiers in both ear-EEG and scalp-EEG ($p < 0.05$), except for the SVM classifier in the scalp-EEG method ($p = 0.12$). When comparing our method with the FBCSP + SVM and EEG + ShallowNet methods, the results show that both approaches are significantly inferior to the TS + MLELM method ($p < 0.01$).

In the comparison of the TS + MLELM and COV + MLELM methods, the result shows that the TS feature extraction method gives a higher mean classification accuracy than the COV method in both ear-EEG and scalp-EEG. However, this is not statistically significant ($p = 0.15$ for ear-EEG and $p = 0.13$ for scalp-EEG).

### 3.4. Comparison of classification results between each session of the experiment

To investigate the effect of training on a user's performance in the speech-imagery task, we compare the classification results from all six sessions of all subjects (figure 10). The results of the statistical test show that there is no notable improvement in results between the first and sixth sessions; in fact, there is no significant change in the results of any pair of sessions ($p > 0.2$). Furthermore, the result also shows no

**Table 2.** Comparision of mean ± std and max ÷ min of the accuracy (%) averaged across all subjects between different methods.

| Method | FBCSP + SVM | TS + LDA | TS + SVM | TS + RVM | TS + ELM | EEG + ShallowNet | COV + MLELM | TS + MLELM |
|---|---|---|---|---|---|---|---|---|
| Ear-EEG | 29.3 ± 6.2<br>36.8 ÷ 22.5 | 28.6 ± 4.9<br>36.7 ÷ 21.5 | 31.9 ± 4.5<br>38.4 ÷ 24.0 | 29.4 ± 3.3<br>34.3 ÷ 24.1 | 32.1 ± 3.6<br>37.8 ÷ 27.1 | 30.54 ± 7.2<br>40.8 ÷ 24.8 | 35.1 ± 3.2<br>42.4 ÷ 31.2 | 37.3 ± 3.2<br>41.9 ÷ 31.9 |
| Scalp-EEG | 32.9 ± 7.7<br>42.1 ÷ 25.4 | 34.2 ± 6.6<br>46.8 ÷ 26.2 | 36.5 ± 7.3<br>50.1 ÷ 28.3 | 33.5 ± 5.6<br>44.0 ÷ 27.2 | 30.6 ± 3.3<br>36.1 ÷ 27.0 | 32.2 ± 10.7<br>46.3 ÷ 20.1 | 37.4 ± 4.6<br>47.2 ÷ 31.8 | 41.7 ± 6.3<br>53.8 ÷ 35.3 |

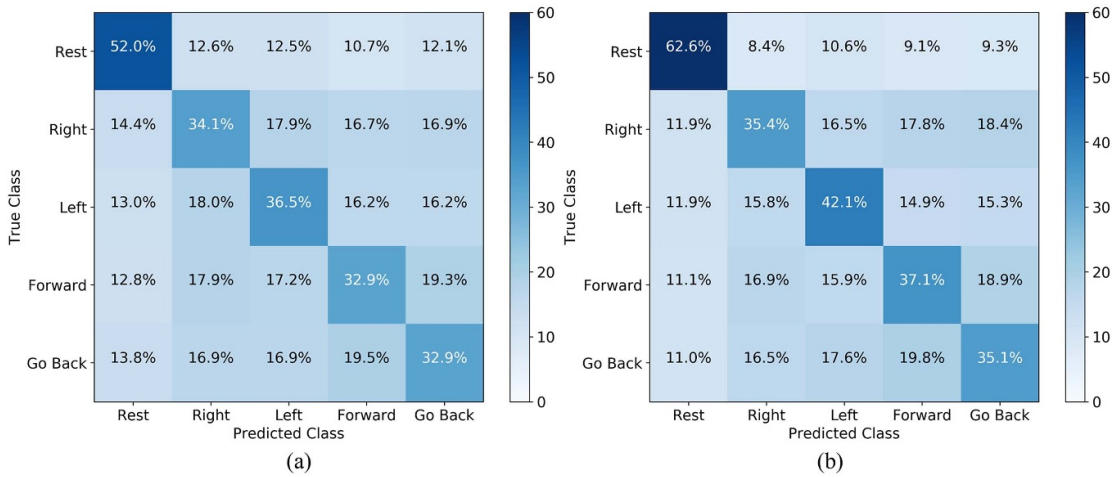


**Figure 9.** Confusion matrix of the results from the TS + MLELM method using (a) ear-EEG and (b) scalp-EEG averaged across all subjects. The color scale indicates the percentage of samples predicted as each class.

significant change between the first and second sessions of the experiment in a day.

## 4. Discussion

### 4.1. Result discussion

The main objective of this study is to examine the performance of ear-EEG in a speech-imagery-based BCI system, and the results from this study show that the performance of the ear-EEG is not inferior to that of the scalp-EEG in most of the subjects. This suggests that ear-EEG has great potential as an alternative EEG acquisition method in speech-imagery-based BCIs.

In an attempt to improve the performance of the ear-EEG, we have performed the EtoS method. However, the results indicate that the current approach to ear-to-scalp feature mapping does not significantly improve the performance of the ear-EEG. Perhaps, a better feature-mapping model has to be developed or more data are needed to properly train the model to make this method work. More studies are needed to address this issue.

When examining the confusion matrix from both ear-EEG and scalp-EEG, we first find that the 'Rest' class has the highest true positive rate. This might be explained by its distinct patterns in the neural activities when compared to the other four speech-imagery tasks. The results also suggest that the 'Right' speech command has the weakest activity in EEG compared to the other speech commands, which causes it to be misclassified as the 'Rest' class most frequently. Finally, we find that words with the same number of syllables are misclassified as each other most frequently. This supports the idea that the number of syllables in a speech command affects the pattern in EEG during the speech-imagery process. However, more extensive experiments are needed to confirm this hypothesis.

One interesting thing that should also be discussed is the poor performance of the ShallowNet method in our data. A possible explanation is that the features that are based on signal amplitude might be dominant in speech-imagery tasks. As previously pointed out in [35], ShallowNet extracts log bandpower as features, which might make it less robust for such BCI paradigms. Furthermore, the number of training samples used to evaluate the system in this study (225 samples, 45 samples for each class) is possibly too low for the ShallowNet model to be appropriately trained.

Finally, the comparisons between the classification results from each session indicate that training does not affect a user's performance in the speech-imagery-based BCI system, which supports the idea that the speech-imagery task is an intuitive mental task that users can perform without any excessive training. This makes the speech-imagery-based BCI system suitable for daily-life use. In addition, the insignificant change between the results of the first and second sessions of each day implies that there is no sign of user fatigue from the experiment.

### 4.2. Neural activity in EEG during speech imagery

The data visualization shows that neural activities in EEG during speech-imagery tasks can be mainly observed in the brain areas that are associated with speech and language. The results from both data visualization and feature analysis indicate that these activities are dominant at high frequency such as in the gamma band, and show the least activity in the theta band.

Previous studies have attributed Broca's area to language production and Wernicke's area to language comprehension. This explains our time-frequency plot for scalp-EEG, where Broca's area shows activity at the onset of trials (i.e. subjects start to imagine a speech in their head). It is also understandable

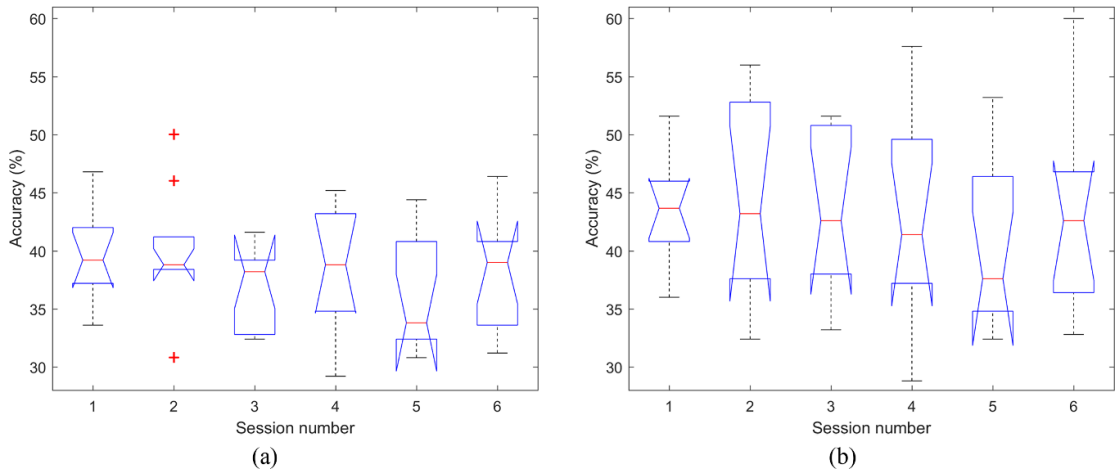


**Figure 10.** Comparison of the classification results (% accuracy) between each session of (a) ear-EEG and (b) scalp-EEG setting. On each box plot, the red central mark is the median, the blue edges are the 25th and 75th percentiles, the black whiskers represent the most extreme data points and the red crosses are outliers. According to the *t*-tests, there are no significant differences between the results from any pair of sessions.

that Broca's area shows high spectral activity for all speech-imagery tasks.

Our spectral analysis of subject S04 suggests that the functions of Wernicke's area may not be exclusive to the semantic aspects of speech in language processing. If Wernicke's area primarily contributes to the semantic comprehension of speech, we expect the single words: 'Left,' 'Right' and 'Forward' to elicit similar responses in this region. In contrast, our analysis shows more similarity between 'Forward' and 'Go back,' where 'Go back,' a phrase made of two words with discrete meanings, should have shown some differences in the activities in Wernicke's area. The study in [36] proposes that Wernicke's area contributes to phoneme perception. This might explain why speech imagery of long speeches that contain a greater number of phonemes and have an interval between two syllables (or words, in the case of 'Go back') displays similar activities in Wernicke's area.

In addition to phoneme perception, the relation between Wernicke's area and cognitive prediction might also contribute to the neural activity in Wernicke's area observed in this study. According to [37], Wernicke's area shows activity in response when making predictions. In our experiment, subjects were aware of what speech to imagine. Activations in Wernicke's area might have been the result of subjects making predictions on what speech to imagine. Thus, it would be interesting to observe the neural activity in Wernicke's area during speech-imagery tasks in an experimental setting that does not give any prior knowledge to the subjects on the corpus of speech-imagery tasks and compare it with the data from this current study.

Another interesting observation from figure 7 is that the time-frequency response of the left channels in ear-EEG and the temporal channels closely resembles the response in Wernicke's area. This suggests that ear-EEG obtained during the speech-imagery tasks in this study is mainly influenced by the activity from Wernicke's area.

We also observe delayed responses in Wernicke's area from long speech imagery in the time-frequency analysis in comparison to short speech. We believe there are two possible explanations for this observation. One reason might be due to our experimental protocol. Subjects were given a loading bar during which they were to pronounce the word in a stretched-out way. While subjects had no problem in this experiment protocol for the short speech, the same might not be said for the long speech. Because the commands used in long speech-imagery tasks are bisyllabic, subjects might have stressed one syllable over the other (in the same way as overt speech production). When we questioned subjects regarding this issue, we found that most of the subjects indeed focused more on the imagery of the second syllable. This might have caused higher activation in EEG from the second syllable compared to the first syllable, which is shown as a delayed response in the time-frequency analysis in Wernicke's area from long-speech-imagery tasks. Another possible explanation is the unclear division between the two syllables. With a single loading bar, subjects might have experienced difficulty in maintaining a consistent rhythm in long speech imagery between trials. Due to different timings between each trial, the slight pause in between the syllables may have been different each time. This might have caused lower activity in earlier parts of speech imagery when we average the trials for the analysis, which in turn results in delayed activity. Furthermore, we believe that the increased period of low neural activity at the beginning of the epoch from the delayed response in Wernicke's area for long speech commands causes the PSD values in this area to be lower when taking the whole EEG epoch into

the calculation, which consequently results in low *F*-values when compared to the 'Rest' class as shown in figure 6.

Additionally, in data visualization, the 'Right' class appears to have the weakest response to speech imagery. One possible explanation is that although the subjects are fluent in English, the pronunciation of the letter 'R' does not exist in the subjects' native language. Therefore, the speech imagery of the word 'Right' might not be executed as well as the other commands, hence the low level of neural activity. We believe that to properly examine the cognitive mechanisms of speech imagery, more extensive studies are required to examine the differences in the pattern in neural activity during speech-imagery tasks between different speech commands, preferably by using better brain-monitoring methods such as fMRI.

It should be emphasized that the above discussions on specific features in the brain patterns during speech-imagery tasks are based on the data visualizations of subject S04 and more studies are needed to make general conclusions on the brain activity during speech imagery.

### 4.3. Choosing the right speech commands

From the results, we can see that choosing the speech commands for speech-imagery tasks is one of the most important things that could affect the performance of the BCI system. In this work, our choices are associated with directions, which could be used in a wide range of applications, such as controlling a wheelchair or a drone. However, it would not be wise to select the words based only on their meaning. Words chosen for the speech-imagery-based BCI should be easily distinguishable in terms of EEG features while retaining their meaning to the specific commands. We believe that more studies are needed to address this matter to find a set of speech commands that optimizes the performance of speech-imagery-based BCI.

We previously discussed and hypothesized that the reason the 'Right' speech-imagery task shows the weakest activity in EEG might be that the pronunciation of the letter 'R' does not exist in the subjects' native language. Following up on this, it would be interesting to see an experiment comparing the performance of speech-imagery-based BCI systems using words from different languages with the same meaning.

### 4.4. Remarks on the ear-EEG acquisition tool

The wearable ear-EEG acquisition tool developed in this study is proven to be successful in acquiring a meaningful signal from the speech-imagery tasks. However, some issues need to be discussed to further improve the equipment. The results from section 3.2.3 show that using only the left channels is enough to obtain meaningful EEG data during the speech-imagery tasks and using the EEG acquired from the right channels does not significantly improve the classification accuracy of the system. This supports the hypothesis that the speech-imagery-related brain activities are dominant in the left hemisphere and that the left channels of the ear-EEG can pick up those signals. However, the reference channel is located on the right side of the equipment, which might cause the signal from the right channels to be weaker due to their proximity. More experiments on the different channel setups on the ear-EEG are needed to confirm the hypothesis. If the system shows a feasible result even when all channels, including the ground and reference channels, are located around the left ear, the equipment could be redesigned to cover only the area around the left ear. This may make the equipment more discreet and comfortable compared to the current design.

Furthermore, we found a problem with the design of the equipment for participants who wear glasses. Since the equipment covers the area around the user's ears, it is uncomfortable to wear glasses together with the ear-EEG device. This problem could be solved by adding a slot on the equipment frame that can be used to attach the glasses' temples to the equipment, or by redesigning the equipment itself in the shape of glasses with sensors located on the temples.

### 4.5. Future work

The next step of this work is to test the system in an online experiment. Here, we only conducted the experiment in an offline manner and evaluated the system by using cross-validation on the entire data to compare the performance between two EEG-acquisition methods. Because of the process of the cross-validation method, the models, including the classifier, the feature selector, and the Riemannian tangent space projector, were different in each iteration of the cross-validation. We also treated the data from each session of a subject separately. In a real-world setting, the cross-validation method will be performed on the data acquired from an offline experiment to find the optimized hyperparameters. The final models will then be trained using the entire data with the optimized hyperparameters before they are used (or tested) in a real-world setting.

Despite the fact that the speech-imagery-based BCI has advantages over other types of BCI, especially in a daily-life setting, it is not yet ready to be used in real-life applications, primarily because of the classification accuracy. According to a survey conducted on 61 people with ALS in [38], the majority of participants preferred command classification accuracy of at least 90%, which unfortunately could not be achieved by the current development of speech-imagery-based BCI. Further improvements in the classification accuracy could be achieved by developing a more powerful algorithm in the data processing, feature extraction method, and classification model.

The system also requires improvements in other aspects. First, most of the current speech-imagery-based BCI studies, including ours, were conducted in laboratories that minimize noises from the environment and the data were acquired while subjects were sitting still. In a real-life setting, there are a lot more noises and EEG artifacts from the constant changes in the environment and the movements of the user. Therefore, noise-canceling and EEG artifact removal methods are required to make the system work efficiently. Second, because EEGs are non-stationary biosignals that could vary over time, environment, and the condition of the human body, the system models require calibration every time before each usage. This issue can be alleviated by using a generic model [39] or transfer learning (TL) techniques. TL is a method that improves generalizability in machine learning models by utilizing knowledge from a source domain to improve the learning performance of a target domain [40]. Recent studies showed that TL techniques can improve the performance of models in speech-imagery-based BCI systems on both a within-subject and inter-subject basis [40, 41].

In addition, since this study was conducted only on healthy subjects, it is necessary to repeat the experiment to confirm that the same results hold for patients who suffer from LIS or ALS. Furthermore, it has been shown in MI research that brain signals from an attempted movement are more similar to signals from the actual movement than those from imagined movement [42], possibly due to motor-inhibitory mechanisms that occur during the MI tasks [43]. MI-based BCI systems that use the attempted movement task also outperform systems that use the MI task [44]. Hence, it would be interesting to conduct a study to see a comparison between the brain signals obtained from the actual, attempted, and imagined speech and their respective performances when they are used in a BCI system.

## 5. Conclusion

In this study, we propose a speech-imagery-based BCI system using ear-EEG as the data acquisition method with the ultimate goal to construct a good framework for daily-life BCI. The proposed system uses the Riemannian tangent space projections of EEG covariance matrices as input features with an MLELM to classify the data. From the data analysis, we find some evidence indicating that the brain activities in Broca's and Wernicke's areas in the gamma frequency band are dominant during the speech-imagery tasks. The results from the multi-class speech imagery experiment show that although scalp-EEG gives a slightly higher accuracy averaged across all subjects, the classification result from ear-EEG is not significantly different from scalp-EEG in seven out of ten subjects. Moreover, mapping the ear-EEG features into the scalp-EEG feature space using an ELM model does not significantly improve the classification accuracy of the system.

Overall, the results from this study show that the ear-EEG acquisition method has great potential to be used as a more convenient and discreet alternative to conventional scalp-EEG for speech-imagery-based BCI systems. It is recommended that future studies on speech-imagery-based BCIs should develop more powerful data processing and machine learning techniques to increase the classification accuracy before using them in real-life applications.

## Acknowledgments

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korea Government (MSIT) under Grant No. 2017-0-00432.

## ORCID iDs

Netiwit Kaongoen https://orcid.org/0000-0002-5935-9662
Jaehoon Choi https://orcid.org/0000-0002-6074-711X
Sungho Jo https://orcid.org/0000-0002-7618-362X

## References

[1] Leuthardt E C, Schalk G, Wolpaw J R, Ojemann J G and Moran D W 2004 A brain–computer interface using electrocorticographic signals in humans *J. Neural. Eng.* **1** 63
[2] Pandarinath C, Nuyujukian P, Blabe C H, Sorice B L, Saab J, Willett F R, Hochberg L R, Shenoy K V and Henderson J M 2017 High performance communication by people with paralysis using an intracortical brain-computer interface *Elife* **6** e18554
[3] Kaongoen N, Yu M and Jo S 2020 Two-factor authentication system using p300 response to a sequence of human photographs *IEEE Trans. Syst. Man Cybern.* **50** 1178–85
[4] Luo A and Sullivan T J 2010 A user-friendly SSVEP-based brain–computer interface using a time-domain classifier *J. Neural. Eng.* **7** 026010
[5] Choi J W, Kim B H, Huh S and Jo S 2020 Observing actions through immersive virtual reality enhances motor imagery training *IEEE Trans. Neural. Syst. Rehabil. Eng.* **28** 1614–22
[6] Fujimaki N, Takeuchi F, Kobayashi T, Kuriki S and Hasuo S 1994 Event-related potentials in silent speech *Brain Topogr.* **6** 259–67
[7] DaSalla C S, Kambara H, Sato M and Koike Y 2009 Single-trial classification of vowel speech imagery using common spatial patterns *Neural Netw.* **22** 1334–9
[8] Matsumoto M and Hori J 2014 Classification of silent speech using support vector machine and relevance vector machine *Appl. Soft Comput.* **20** 95–102
[9] Deng S, Srinivasan R, Lappas T and D'Zmura M 2010 EEG classification of imagined syllable rhythm using Hilbert spectrum methods *J. Neural. Eng.* **7** 046006
[10] Leuthardt E C, Gaona C, Sharma M, Szrama N, Roland J, Freudenberg Z, Solis J, Breshears J and Schalk G 2011 Using the electrocorticographic speech network to control a brain–computer interface in humans *J. Neural. Eng.* **8** 036004
[11] Pei X, Barbour D L, Leuthardt E C and Schalk G 2011 Decoding vowels and consonants in spoken and imagined

words using electrocorticographic signals in humans *J. Neural. Eng.* **8** 046028
[12] Martin S, Brunner P, Iturrate I, Millán J D, Schalk G, Knight R T and Pasley B N 2016 Word pair classification during imagined speech using direct brain recordings *Sci. Rep.* **6** 25803
[13] Nguyen C H, Karavas G K and Artemiadis P 2017 Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features *J. Neural. Eng.* **15** 016002
[14] Qureshi M N, Min B, Park H J, Cho D, Choi W and Lee B 2017 Multiclass classification of word imagination speech with hybrid connectivity features *IEEE Trans. Biomed. Eng.* **65** 2168–77
[15] García-Salinas J S, Villaseñor-Pineda L, Reyes-García C A and Torres-García A A 2019 Transfer learning in imagined speech EEG-based BCIs *Biomed. Signal Process. Control* **50** 151–7
[16] Lee S H, Lee M, Jeong J H and Lee S W 2019 Towards an EEG-based intuitive BCI communication system using imagined speech and visual imagery *2019 IEEE Int. Conf. on Systems, Man and Cybernetics (SMC) (6 October)* (IEEE) pp 4409–14
[17] Debener S, Emkes R, De Vos M and Bleichner M 2015 Unobtrusive ambulatory EEG using a smartphone and flexible printed electrodes around the ear *Sci. Rep.* **5** 16743
[18] Mikkelsen K B, Kappel S L, Mandic D P and Kidmose P 2015 EEG recorded from the ear: characterizing the ear-EEG method *Front. Neurosci.* **9** 438
[19] Kaongoen N and Jo S 2020 An ear-EEG-based brain–computer interface using concentration level for control *2020 8th Int. Winter Conf. on Brain–Computer Interface (BCI) (26 February)* (IEEE) pp 1–4
[20] Fiedler L, Wöstmann M, Graversen C, Brandmeyer A, Lunner T and Obleser J 2017 Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech *J. Neural. Eng.* **14** 036020
[21] Bleichner M G, Mirkovic B and Debener S 2016 Identifying auditory attention with ear-EEG: cEEGrid versus high-density cap-EEG comparison *J. Neural. Eng.* **13** 066004
[22] Mikkelsen K B, Villadsen D B, Otto M and Kidmose P 2017 Automatic sleep staging using ear-EEG *Biomed. Eng. Online* **16** 111
[23] Looney D, Kidmose P and Mandic D P 2014 Ear-EEG: user-centered and wearable BCI *Brain–Computer Interface Research* (Springer: Berlin) pp 41–50
[24] Kidmose P, Looney D and Mandic D P 2012 Auditory evoked responses from Ear-EEG recordings *2012 Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society (28 August)* (IEEE) pp 586–9
[25] Kaongoen N and Jo S 2018 An auditory P300-based brain-computer interface using Ear-EEG *2018 6th Int. Conf. on Brain-Computer Interface (BCI) (15 January)* (IEEE) pp 1–4
[26] Friedman L, Kenny J T, Wise A L, Wu D, Stuve T A, Miller D A, Jesberger J A and Lewin J S 1998 Brain activation during silent word generation evaluated with functional MRI *Brain Lang.* **64** 231–56
[27] Binder J R 2015 The Wernicke area: modern evidence and a reinterpretation *Neurology* **85** 2170–5
[28] Koessler L, Maillard L, Benhadid A, Vignal J P, Felblinger J, Vespignani H and Braun M 2009 Automated cortical projection of EEG sensors: anatomical correlation via the international 10–10 system *Neuroimage* **46** 64–72
[29] Gaur P, Pachori R B, Wang H and Prasad G 2018 A multi-class EEG-based BCI classification using multivariate empirical mode decomposition based filtering and Riemannian geometry *Expert Syst. Appl.* **95** 201–11
[30] Huang G B, Zhu Q Y and Siew C K 2006 Extreme learning machine: theory and applications *Neurocomputing* **70** 489–501
[31] Ding S, Zhang N, Xu X, Guo L and Zhang J 2015 Deep extreme learning machine and its application in EEG classification *Math. Probl. Eng.* **2015** 129021
[32] Tipping M E 2000 The relevance vector machine *Advances in Neural Information Processing Systems 12* Eds. (Cambridge, MA: MIT Press) pp 652–8
[33] Ang K K, Chin Z Y, Zhang H and Guan C 2008 Filter bank common spatial pattern (FBCSP) in brain–computer interface *2008 IEEE Int. Joint Conf. on Neural Networks (IEEE World Congress on Computational Intelligence) (1 June)* (IEEE) pp 2390–7
[34] Schirrmeister R T, Springenberg J T, Fiederer L D, Glasstetter M, Eggensperger K, Tangermann M, Hutter F, Burgard W and Ball T 2017 Deep learning with convolutional neural networks for EEG decoding and visualization *Hum. Brain Mapp.* **38** 5391–420
[35] Lawhern V J, Solon A J, Waytowich N R, Gordon S M, Hung C P and Lance B J 2018 EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces *J. Neural. Eng.* **15** 056013
[36] Binder J R 2017 Current controversies on Wernicke's area and its role in language *Curr. Neurol. Neurosci. Rep.* **17** 58
[37] Bischoff-Grethe A, Proper S M, Mao H, Daniels K A and Berns G S 2000 Conscious and unconscious processing of nonverbal predictability in Wernicke's area *J. Neurosci.* **20** 1975–81
[38] Huggins J E, Wren P A and Gruis K L 2011 What would brain–computer interface users want? Opinions and priorities of potential users with amyotrophic lateral sclerosis *Amyotroph. Lateral Scler.* **12** 318–24
[39] Jin J, Sellers E W, Zhang Y, Daly I, Wang X and Cichocki A 2013 Whether generic model works for rapid ERP-based BCI calibration *J. Neurosci. Methods* **212** 94–9
[40] Cooney C, Folli R and Coyle D 2019 Optimizing layers improves CNN generalization and transfer learning for imagined speech decoding from EEG *2019 IEEE Int. Conf. on Systems, Man and Cybernetics (SMC) (6 October)* (IEEE) pp 1311–6
[41] Tamm M O, Muhammad Y and Muhammad N 2020 Classification of vowels from imagined speech with convolutional neural networks *Computers* **9** 46
[42] Bruurmijn M L, Pereboom I P, Vansteensel M J, Raemaekers M A and Ramsey N F 2017 Preservation of hand movement representation in the sensorimotor areas of amputees *Brain* **140** 3166–78
[43] Guillot A, Di Rienzo F, MacIntyre T, Moran A and Collet C 2012 Imagining is not doing but involves specific motor commands: a review of experimental data related to motor inhibition *Front. Hum. Neurosci.* **6** 247
[44] Blokland Y, Vlek R, Karaman B, Özin F, Thijssen D, Eijsvogels T, Colier W, Floor-Westerdijk M, Bruhn J and Farquhar J 2012 Detection of event-related desynchronization during attempted and imagined movements in tetraplegics for brain switch control *2012 Annual Int. Conf. IEEE Engineering in Medicine and Biology Society (28 August)* (IEEE) pp 3967–9