A Novel Onset Detection Technique for Brain-Computer Interfaces using Sound-production Related Cognitive Tasks in Simulated-online System

YoungJae Song[1], Francisco Sepulveda[1]

[1] BCI and Neural Engineering Group – School of Computer Science and Electronic Engineering, University of Essex, United Kingdom

E-mail: syoungb@essex.ac.uk

Abstract

Objective. Self-paced EEG-based BCIs (spBCIs) have traditionally been avoided due to two sources of uncertainty: 1) precisely when an intentional command is sent by the brain, i.e., the command onset detection problem, and 2) how different the intentional command is when compared to non-specific (or idle) states. Performance evaluation is also a problem and there are no suitable standard metrics available. In this paper we attempted to tackle these issues.

Approach. Self-paced covert sound-production cognitive tasks (i.e., high pitch and siren-like sounds) were used to distinguish between intentional commands (IC) and idle states. The IC states were chosen for their ease of execution and negligible overlap with common cognitive states. Band power and a digital wavelet transform were used for feature extraction, and the Davies-Bouldin index was used for feature selection. Classification was performed using LDA.

Main results. Performance was evaluated under offline and simulated-online conditions. For the latter, a performance score called true-false-positive (TFP) rate, ranging from 0 (poor) to 100 (perfect), was created to take into account both classification performance and onset timing errors. Averaging the results from the best performing IC task for all seven participants, an 77.7% true-positive rate was achieved in offline testing. For simulated-online analysis the best IC average TFP score was 76.67% (87.61% true-positive rate, 4.05% false-positive rate).

Significance. Results were promising when compared to previous IC onset detection studies using motor imagery, in which best true-positive rates were reported as 72.0% and 79.7%, and which, crucially, did not take timing errors into account. Moreover, based on our literature review, there is no previous covert sound-production onset detection system for spBCIs. Results showed that the proposed onset detection technique and TFP performance metric have good potential for use in spBCIs.

Keywords: Brain-Computer interface; Onset detection; Covert sound-production; Self-paced BCI; Asynchronous BCI; EEG; True-False-Positive Rate Score

I. Introduction

There are several different manners of categorising BCIs. Amongst these definitions, BCIs can also be categorised as cue-based (synchronous) or self-paced (asynchronous) systems [1]. Cue-based (CB-BCI) and self-paced BCIs (SP-BCI) systems have different approaches to interact with users. CB-BCI systems tell the users when to start and stop a relevant brain activity task that will lead a command to the machine. CB-BCIs include P300 and SSVEP systems as well as those based on cue-based cognitive tasks. These approaches force the users to keep their mental focus and/or gaze on the computer interface (i.e., typically a computer-controlled visual or auditory stimulus [2-4]) which is not only very unnatural to users, but also leads to loss of both user autonomy and the ability to have a rich interaction with their environment. The majority of current EEG-based BCI systems are CB-BCIs. The advantage of CB-BCIs is that they give better classification rates and easier analysis than SP-BCIs as the machine is not required to determine the time location of relevant events; i.e., the machine only needs to determine what the user intended to do, not when a relevant mental state is present. This is crucial as the brain is constantly multitasking – at some level – making it difficult to determine when exactly the user intended to communicate with the machine by means of brain signals alone.

SP-BCIs, on the other hand, analyse user's brain signals continuously without a specific computer-controlled stimulus [1]. The users control the timing of the BCI system by intentionally performing a specific cognitive task when it suits them [3], thus providing increased autonomy, flexibility, and interaction with the environment (including the people therein, of course). For this reason, SP-BCIs are more suitable than CB-BCIs for the ultimate aim of transferring BCIs from laboratory settings towards real-world use.

However, there are great challenges in SP-BCIs [3]. Due to the system's lack of knowledge about the precise time location of user command, SP-BCIs need to continuously analyse the ongoing brain activity in order to classify between intentional-control (IC) and non-control (NC) states (also called non-specific or null states). NC states can be any states besides IC states (e.g., idle, daydreaming, other mental activities, irrelevant evoked responses, etc.) [4]. One way to distinguish between IC and NC is to use a classifier that treats NC and IC as just different states in the same classification task. For example, a five-output classifier can include NC states as one of the five output classes. However, given the brain's constant multitasking, this approach – herein called a 'lumped' approach – will lead to a high false-positive rate for the IC states (and thus a high false-negative rate for the NC states) and to large timing errors in IC detection. Hence, an alternative approach is needed, namely, separating the '*when*' classification task (herein called IC **onset detection**) from the '*what/which*' classification stage. This simplifies the problem and leads to reduced timing errors and lower NC misclassification rates, although the approach is not entirely issue-free. Three important factors to consider when using IC onset detection are: a) the high asymmetry in data set sizes (i.e., there will be much more NC than IC data), b) determining what is an acceptable timing error, and c) the fact that the overall system's performance is not the onset true positive rate (OS_tpr) added to or averaged with the classification rate for the separate IC states (ICstates_tpr). Instead, the overall classification performance will be determined by the product of these two quantities, i.e., OS_tpr x ICstates_tpr, making the performance more sensitive to OS_tpr than in the lumped approach described above. This highlight the need to improve the OS_tpr as much as possible without sacrificing timing accuracy.

In this paper, sound-production related cognitive tasks have been proposed for the onset detection method. Based on our thorough literature review (up to 2016), none of the work on onset detection or self-paced BCIs systems used speech or sound-production related cognitive tasks. They mostly used motor imagery (e.g., [3-6]). In addition, all the speech related EEG based BCI studies using different syllables (or syllables/vowels) that we found were focused on discrimination between various tasks, not on onset detection (i.e., idle vs. intentional state) and were cue-based approaches, not self-paced (e.g., [7-10]), and some were ECoG studies [11, 12]. This is the main novelty in our study: discriminating between sound-production related tasks and idle (or non-specific) states for onset detection, which led to competitive results compared to systems based on typical motor-imagery tasks [5, 13]. We also introduce a novel score system for evaluating self-paced BCI performance.

Sound-production related tasks are very intuitive for the vast majority of people as we almost constantly 'speak' internally or imagine many words in normal life. This is also a big advantage for people with severe motor disabilities, an important target population for BCIs. The challenge, however, is to reduce chances of IC false positives, which can be addressed by choosing cognitive tasks that do not significantly overlap with other common, spontaneous and frequent cognitive states [14]. Using specific words/syllables/letters for onset detection would likely increase both onset false-positives as well as task-related false negatives due to large overlap with the continuous internal speech in normal thought processes. For this reason, we have chosen imagining a high tone or siren-like sound production tasks (with covert and inhibited-overt execution, for comparison purposes, respectively) as onset switches, both of which are unlikely to overlap with normal thought processes. In addition, our chosen tasks are easy to produce and control voluntarily and there is no dependence on the subjects' mother-language. We have tested onset detection in offline and simulated online scenarios as a prototype towards practical online system.

II. Methodology

A. Sound-production related Tasks and Idle State Definition

In this experiment, there were two different mental tasks for the onset switch, and two modes for each task. Firstly, the modes are separated as in inhibited overt (*IO*) and covert (*C*) sound production. Secondly, high tone (*High*) and siren-like (*Siren*) sound production mental tasks were tested. For the non-control state, idle (*Idle*), i.e., non-specific states were also

recorded (to avoid confusion, the term '*idle*' alone will be used in the remaining parts of this paper). The start and duration of the tasks was controlled spontaneously by the user (assisted by a specially designed time-keeping interface, described below). To minimise artefacts generated from muscle signals, participants were instructed to avoid any unnecessary body movement, but they were still allowed to blink or move their eyes when needed (the artefact rejection methods are explained later in this paper).

In more detail, the states were defined as follows:

- **Inhibited overt sound-production Tasks:**

  Inhibited overt sound-production is different from our normal overt sound-production. Aside from the cognitive effort, it will involve tensioning of the vocal cords but there is no actual sound production that can clearly be heard.

  ***Inhibited overt high tone production (IO_High)*:** participants were instructed to produce an 'um' sound effort with a high pitch that they can comfortably produce for a couple of seconds, but high enough that they think it is an unusual tone and not something they would imagine in a normal situation.

  ***Inhibited overt siren-like sound production (IO_Siren)*:** the siren-like sound effort was defined as 'wee-woo wee-woo'. 'Wee' syllable denotes high notes, whereas 'woo' expresses low pitch. Participants were instructed to produce this sound effort for a couple of seconds.

- **Covert sound-production Tasks:**

  Covert sound-production was a pure imagination process. Thus, there should be no tensioning of any organs related to sound-production. Participants were instructed to imagine making the 'sound', which of necessity included imagining hearing the sound (auditory imagery / auditory recall). Auditory imagery refers to mental imagery in sound perception without an actual external auditory stimuli [15]. In terms of functional neuroanatomy the processes involved in covert speech have not be fully elucidated, but it is known that it involves the auditory cortex (around Brodmann areas 41, 42 and partially 22 [16]) and, for speech-related imagery, Wernicke's area (Brodmann area 22). Also, the auditory system has been shown to play an important role in overt speech production by giving internal feedback [17], it is possible that a similar role is played in covert sound-production.

  ***Covert high tone production (C_High)*:** Imagining high tone production (as explained above for the *IO_High* task).

  ***Covert siren-like sound production (C_Siren)*:** Participants were instructed to imagine making siren-like sounds in covert mode.

- **Idle state (*Idle*):** This is a non-control or null state. The participants were instructed to not think of any of the above IC task states and to stay calm and relax.

During all above tasks, participants were not allowed to imagine tongue, mouth, lips, or any other body movements to avoid motor-imagery related signals.

B. Experiment Interface Design

While the tasks were controlled spontaneously by the users, it was necessary to provide them with a means to estimate the length of time gone by when executing a task in order to ensure that the IC task lasted sufficiently long to yield enough data to achieve a high classification rate, but not so long that it would lead to such high timing errors as to render the self-paced approach useless. Having in mind the typical task duration in cue-based BCIs, we chose an approximate recommended task duration of 3s, but bear in mind that the user was still free to start and stop the task at any time that suited them, within a 30s window (the maximum duration of each NC state within a trial so that the experiments did not run for an unnecessarily long time).

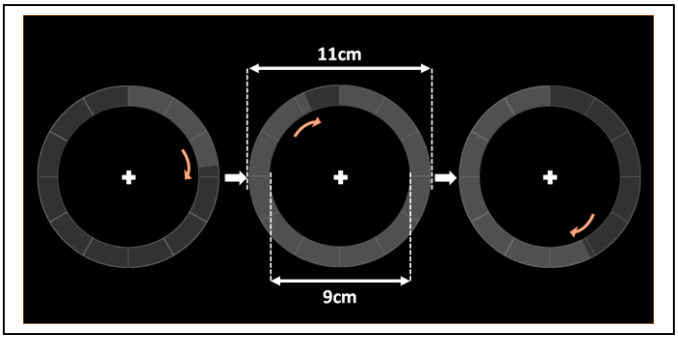

Figure 1. The chosen time-keeping interface design. Users fix their eyes on the central cross and estimate their task time as the light grey progress bar grows clockwise.

To record onset tasks and idle states for simulated-online scenarios (i.e., treating the data trials sequentially rather than as independent random trials), the time tracking interface needs to be suitable for actual online self-paced onset detection systems even during recording of the training. Thus, there were three main functional requirements: a) The interface should minimise visual event-related potentials (VEP). b) The computer must be able to time-stamp events. And, c) as explained above, the user must be able to estimate task duration. To satisfy these requirements, a few different recording interfaces were considered as candidates and the circular progress bar interface shown here was chosen based on the facts of 1) minimum eye movement, 2) minimum ERP generation, and 3) usability (defined as ease of use in this study) from three experienced BCI users (i.e., PhD students in our BCI group). To determine the size of the interface, we considered two literature sources. In [18] competing stimuli located less than 5° of visual angle from the central stimulus were shown to affect SSVEP responses. In [19], similar effects were observed in a P300-

based BCI. As a result, to avoid these proximity issues, the diameter of the interface's inner circle was set to 9cm and the distance between the monitor and participants was set to 50cm. This leads to about 10° of viewing angle. The viewing angle from the fixation cross to any circular moving object was just above 5°. In addition to this, background and objects colour were chosen to be dark achromatic colours to minimise ERPs. As shown in Figure 1, the progress bar in the interface continuously filled with light grey for 12 seconds and then with slightly darker grey (the jump in brightness was small to minimise VEP), followed by light grey again.

C. Experimental Protocol

Participants performed one run for each task, chosen pseudo-randomly to minimise sequence-dependent effects (randomisation between runs). In each run, participants executed the same task 30 times. They knew which task to perform as they were told about the task, by the experimenter, before each run. Task randomisation within a run was unnecessary and undesired in our case as this is only relevant in a multi-task scenario (e.g., motor imagery for left hand vs. right hand vs. feet vs. tongue, etc.). In our case, on the other hand, the intended task-versus-idle scenario is one in which the end-user would execute the same imagery task every time. I.e., in an onset-detection problem it would make no sense to mix the tasks, as this is not what will be happen in online use.

Between each 30-trial run, participants had short breaks (1-3min, as desired). The total experiment time did not exceed one hour beyond electrode cap set up and explanation of the experiment to the participant.

The experiments were done in accordance with the University of Essex Ethics Committee guidelines.

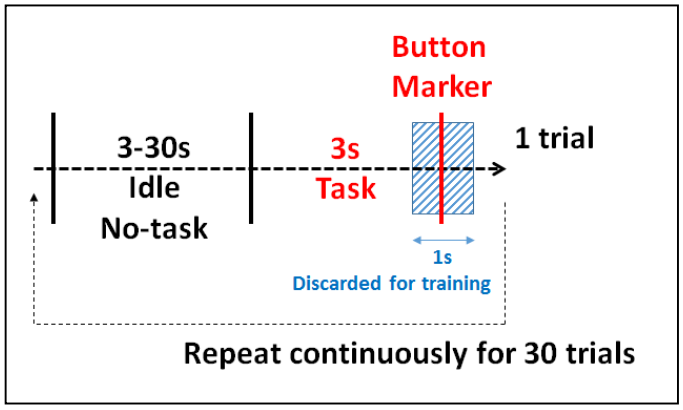

Figure 2. EEG data recording procedure for 1 trial.

Figure 2 represents the recording procedure for one trial. Users were required to stay in the idle state for at least 3s, after which they were free to execute one of the cognitive tasks at any time up to 30s from the beginning of the trial. Immediately after they executed a task for about 3s (aided by the time-keeping interfaces) they were required to press the space key on the keyboard to signal the end of a trial and to provide a time stamp for performance evaluation of the system. The minimum idle state of 3s was chosen to prevent task time-proximity effects in the EEG. On the other hand, the maximum idle state 30s was chosen based on a previous study [14] that explored different ways of time-stamping active states and in which participants were given a window of up to 100s within which to execute the self-paced task. In that study all participants spontaneously executed a given active task within 15s after a trial began. For this experiment, an extra 15s were included to prevent participants from rushing.

During the whole experiment the same key (the space key) was pressed following a self-paced task to prevent class-dependent information from any motor-related signal. In addition, data 0.5s prior to and 0.5s after the space key (shaded area is Figure 2) were discarded from the analysis to avoid motor-imagery related data leading to IC false-positives.

Seven healthy subjects (4 males, 3 females) participated in the experiments. They all had normal or corrected vision and were aged between 22 and 27. Three participants had previous experience with BCIs and two of them had participated in a previous study on covert sound production for onset detection [14]. The other four participants were naïve subjects. Each subject was sat on a medical chair comfortably and a monitor was placed 50cm away from subject's face. A keyboard was placed on their lap to press the space bar for the end-of-trial marker.

D. Offline and Simulated-online Evaluation Definition

In this experiment, the recorded data were analysed in offline and simulated-online scenarios, as follows.

***Offline evaluation*:** The continuously recorded EEG data was segmented into 0.5s time windows without any overlapping. Then, these segments were separated into task and idle states based on the timing protocol shown in Figure 2. If a segmented 0.5s window included both idle and task states, it was discarded, as were the 0.5s before and after the key-press stamp. After segmentation, half of the epochs for each state were randomly selected for training and the other half were used for testing data. The randomisation-training-testing cycle was repeated 20 times. Offline evaluation gives a preliminary idea about how well the system can distinguish active tasks from idle states for onset detection and the results can more easily be compared to other BCI systems. However, the offline evaluation has drawback towards real onset detection system as it ignores sequence effects (such as possible priming, habituation, etc.) of onset tasks.

***Simulated-online evaluation*:** Data segmentation was done as in the offline study, but with two crucial differences: a) no data windows were discarded unless EOG was automatically detected by the system (using the EOG detection algorithm described below), and b) epoch randomisation was not applied in order to preserve the online-like time structure of the data. Instead, the first 15 trials (half of the recorded trials within a run) were used for training; the subsequent 15 trials were used for testing. Data were not discarded in the manner done in the offline approach because in real online situations there is no end-of-trial marker.

E. Data Recording and Signal Pre-Processing

A Biosemi (TM) ActiveTwo system was used with the Actiview software for recording data. 64 electrodes were placed based on 10-10 layout system and 2 reference electrodes were placed on the left and right earlobes. In addition, 1 electrode was setup to detect EOG artefacts. Sampling rate 512 samples/s was chosen to ensure recording up to the high gamma band (100 Hz) based on 3dB-point (half power point) of the equipment bandwidth around 104 Hz. In BCI studies, high gamma waves have not been investigated very often due to increased contamination by muscle artefacts, but previous work by others [20-22] has shown significant high gamma wave activity associated with some language tasks, hence its inclusion here. On the other hand, recording at a higher rate was not necessary as our interest in EMG was only for artefact removal purposes and, further, sampling at a higher rate could have led to increased EMG-related aliasing in the EEG signals.

Continuously recorded EEG data were segmented with 0.5s window length. The data were band-pass filtered (zero-lag Butterworth filter, order 4) with cut-off frequencies at 2 Hz and 100 Hz. Then a notch filter (zero-lag Butterworth filter, order 4) was applied at 49-51 Hz to reduce mains interference. To remove common environmental noise, the averaged of the two earlobe reference channels was subtracted from all 64 scalp channels.

F. EOG Artefact Detection

An EOG channel was placed above the corrugator muscle and was used for EOG detection at the forehead region. Figure 3 illustrates the procedure for automatic EOG detection. A discrete wavelet transform (DWT) with Haar mother wavelet (because it resembles eye blink ocular artefacts [23]) was applied to the EOG channel. The decomposition level, 6, was chosen as it showed satisfactory results in [23, 24]. The pseudo-frequency of the level 6 approximation component was 0-8Hz in our case.

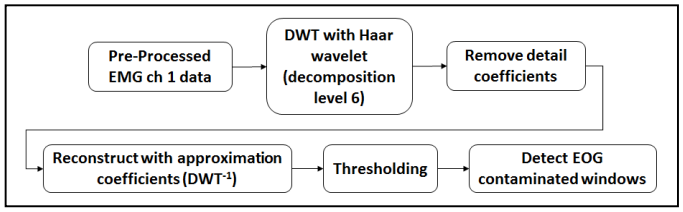

Figure 3. Block diagram of EOG artefact detection method.

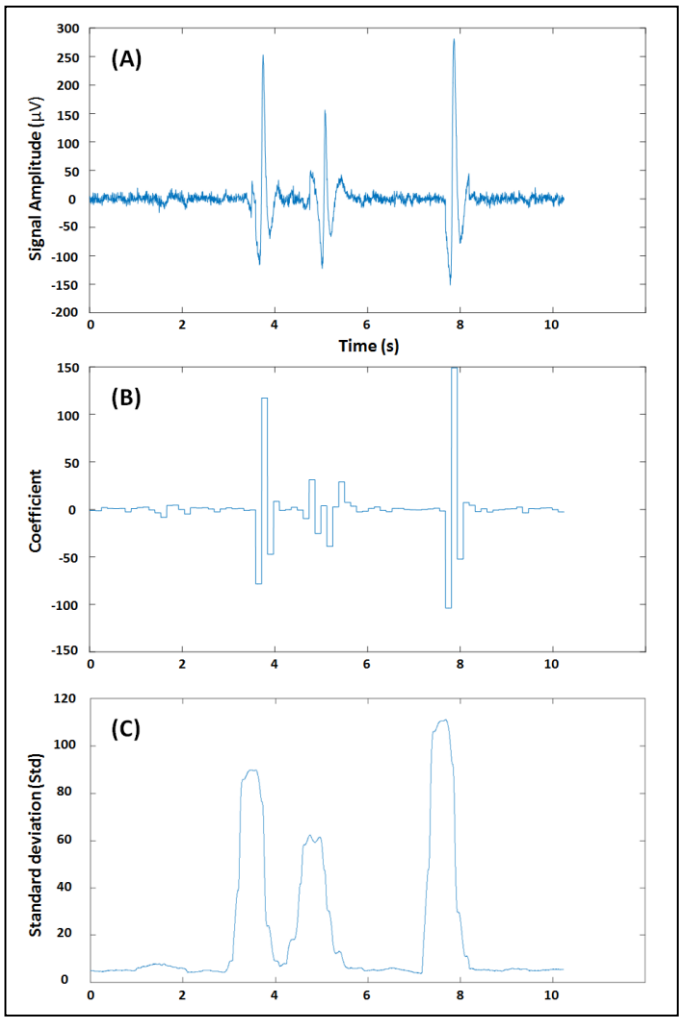

Figure 4. Participant 1's first 10s data (A) Pre-processed EOG channel. (B) EOG artefact detection process applied with wavelet transform. (C) Standard deviation of 0.5s data from (A).

To detect EOG artefacts, two conditions needed to be met: i) a standard deviation (**std**, calculated for each from 0.5s non-overlap window segment) jump by a factor of 3, and ii) using a wavelet coefficient threshold, as follows. If we compare Figure 4A and B, the EOG detection plot (B, based on the wavelet coefficients at decomposition level 6) can be seen to have large rising/falling edges. When the standard deviation (**std**) was found to jump by a factor of 3, the subsequent data were treated as possible EOG artefact candidates. Within the EOG artefact region, the smallest rising/falling step area was chosen as a threshold in order to avoid discarding false EOG positives that may result from applying only the **3std** condition. E.g., in Figure 4B, between 5s and 6s we find a pattern that can be deemed to be border line EOG artefact and, within that region, the smallest step is 20μV. This value was half powered (-3db) and the result was chosen as a threshold. To reduce onset false positives, once the EOG artefact contaminated time-locations are detected, the data for those segments were discarded from further analysis.

G. EEG Feature Extraction

In order to analyse the EEG signal, two different feature extraction methods were used, band power and wavelets. For the band power a Fast Fourier Transform was applied to the pre-processed EEG signals and its power (i.e., the squared FFT) were selected as features from eight different frequency ranges; Freq1: 2-4Hz (Delta), Freq2: 4-8Hz (Theta), Freq3: 8-12Hz (Alpha), Freq4: 12-16Hz (Low Beta), Freq5: 16-20Hz (Beta), Freq6: 20-30 (High Beta), Freq7: 30-42Hz (Low Gamma) and Freq8: 42-100Hz (High Gamma).

The second feature extraction method was the discrete wavelet transform. It offers time-frequency features and performs well with non-stationary brain signals [25]. Pre-processed EEG signals were decomposed and their coefficient vectors from levels 6Approximation, 6Detail, 5D, 4D, 3D and 2D (representing the pseudo frequency bands Wave F1: 2-4Hz, Wave F2: 4-8Hz, Wave F3: 8-16Hz, Wave F4: 16-32Hz, Wave F5: 32-64Hz and Wave F6: 64-100Hz respectively) were calculated and their variances (for dimensionality reduction purposes) were used as features. The mother wavelet 'db2' was chosen because of its simplicity and common use in EEG signal analysis [26-28] (also, in our previous study [14] we found that the choice of wavelet type – db2, coif2, or sym2 – did not significantly affect sound-production related onset detection). While it is possible that an extensive study including various other wavelet types and orders (and, for that matter, other JTFA and non-JTFA approaches) could lead to improved results, our study was meant to focus on the use of covert sound-production in onset detection.

H. Classification

The above feature extraction method produced hundreds of features, i.e., (64ch*7band power + 64ch*5wavelet = 768 features), so feature selection had to be applied to reduce feature set size and class overlap, and to improve computational efficiency. To this end, the Davies-Bouldin index (DBI [29, 30]) was applied. The DBI is a cluster overlap measure. Smaller DBI values indicate better class separation, with lower class overlap and larger distance between classes. Thus, DBI values for each feature were sorted in ascending order and an integer value DBI threshold from 1 to ***N*** was obtained for each participant. The features which had DBI value less than the threshold were selected as a feature set for further analysis. The DBI threshold was chosen as follows.

The DBI threshold was chosen based on the training set's classification result (see Figure 5). Due to the different sizes of the idle and task states (the idle period is much longer than tasks), classification results could be biased towards the idle state (see points DB=1 and 2 in the figure). By increasing the DBI threshold (e.g., from 2 to 3), the task state's true-positive rate increases while the idle true-positive rate decreases. This behaviour continues until the individual TP rate continuously decreases for both idle and task states. However, in every case there is an optimum DBI value at which the overall TP rate is maximised (e.g., at DBI=4 in Figure 5). Thus, the DBI threshold was chosen so that it gave the highest overall true-positive rate for the training data.

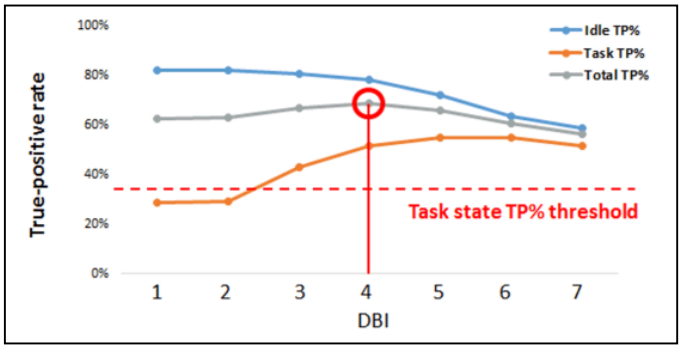

Figure 5. Sample training true-positive rates for idle, task periods, and total performance (from participant 1, inhibited-overt siren task). The horizontal axis shows 7 approximate DBI values for illustration purposes.

After feature selection was performed, Linear Discriminant Analysis (LDA) was applied for classification. LDA was chosen due to its simplicity and low computational power [31] as well due to its widespread use in BCI research. The feature vectors from the feature selection process were used as inputs to the LDA. For the offline analysis, pseudo-randomisation of the choice of training and testing set epochs was done 20 times and results obtained for each randomisation stage.

I. EMG Artefact Handling

A challenge in all BCIs, but more so when the gamma band is included in the analysis, is to ensure that classification results are based on brain signals alone, as much as possible, and are not contaminated by potentially class-dependent EMG. In an EMG artefact BCI survey [32], it was shown that 67.5% of the BCI studies included in the survey did not mention whether they handled EMG artefacts or not and 12.1% did not remove EMG artefacts.

EMG artefacts are particularly important for IC state onset detection as switching from an 'idle' state to an IC state may produce involuntary facial twitches that can produce class-dependent EMG artefacts, especially in frontal area EEG, more so than when switching between various IC states. EMG (and other facial artefacts) must thus be minimised.

Independent component analysis (ICA) and blind source separation by canonical correlation analysis (BSS-CCA) are the two mostly used EMG removal techniques in BCIs. Research papers [33, 34] showed BSS-CCA outperformed ICA and it was more suitable for EMG removal thus BSS-CCA was chosen for this experiment.

CCA measures the linear relationship between two multi-dimensional signals [35]. It can be used to solve BSS problem (proposed in [36]) by taking multi-channel EEG as a first variables and temporally delayed version as a second variables [37]. The threshold of autocorrelation coefficient ρ was chosen as 0.35 based on the study in [38]. If there was no source that has less than the threshold ρ, the last source (from descending

order sort) that has the lowest autocorrelation coefficient was removed.

J. Performance Assessment (True-False-Positive Score)

For the event by event performance evaluation, the true-false difference rate was suggested in [39] for self-paced BCIs. However, there are some issues with this approach. Due to the difficulty in measuring true-negatives during idle state, [39] proposed a false-positive rate as ***'FP/(E+FP)'***, where FP is the number of false-positives and E is the number of task state onset events. This false-positive rate was subtracted from the true-positive rate. However, the number of task events and idle events are independent in self-paced system. Yet, the method in [39] would yield the same score even if two different systems have different lengths for the idle states but have the same amount of false-positives. The system with longer idle periods should yield a higher score as this system makes less frequent IC onset false-positives, and is thus more robust, but that is not what the index in [39] would indicate

Thus, to address the limitations in [39], we propose a new performance evaluation score, called *true-false-positive score (*TFP_{Score}*)*, defined as follows:

$$\mathrm{TFP}_{Score}\,(\%) = \frac{(TP+\alpha)}{(tE+\alpha)} * \left(1 - \frac{(FP+\alpha)}{(iE+\alpha)}\right)^2 * 100 \qquad (1)$$

where TP and FP are the numbers of true-positive and false-positive 500ms-windows, respectively in this study. tE and iE refer the number of IC task onset events and idle events, respectively. 'α' is set to 0.1, which is a very small number chosen merely to avoid division by zero while still minimising effects on the results. To define iE more clearly, the different online system time periods will be defined as follows:

- ***Recording Time*:** Total recording time for a run without any stops and interruptions.
- ***Task Period*:** Total task activation time, i.e., the sum of all task activation periods (from the beginning of task activation to the stop). This variable includes a timing error tolerance region (described below in Results & Discussion). If the experiment is designed to maintain the task activation state until the user receives feedback, then the tolerance region is not included.
- ***Refractory Period*:** Period in which the signal is ignored after the task activation or false-positive, i.e., the machine ignores incoming data while it executes whatever function is required after onset detection.
- ***Idle Period*:** Total idle state period, ***Idle Period = Recording Time - Task Period - Refractory Period***.

iE will be defined as the number of shifting windows that give classification results as Idle (e.g., assuming a non-overlapping window size of 500ms, a 1s idle period gives iE = 2). The behaviour of eq. (1) is shown in Figure 6. Ignoring α for simplicity, we obtain the following behaviours, all of which are correct:

- When FP is zero, TFP will vary with iE, so, everything being equal, longer idle periods will yield higher performance scores.
- By multiplying (1-FP rate) to TP rate, the score is reduced if FP is increased.
- If FP is zero, the score will be near TP and will depend on idle period size.
- **FP rate=top/bottom.** The square power of (1-FP rate) will give more reasonable scores than by removing the power of 2. For example, TP=6 and FP=0 give a TFT score around 60%. The score will be similar when TP=7 and FP=3 in panel A (i.e., with the power of 2). However, without the power of 2 (panel B), a score near 60% would be obtained with TP=7 and FP=7, which does not make sense as a system with TP=6 and FP=0 is clearly much better than one with TP=7 and FP=7. For this reason, the square power was chosen after investigation with many scenarios.

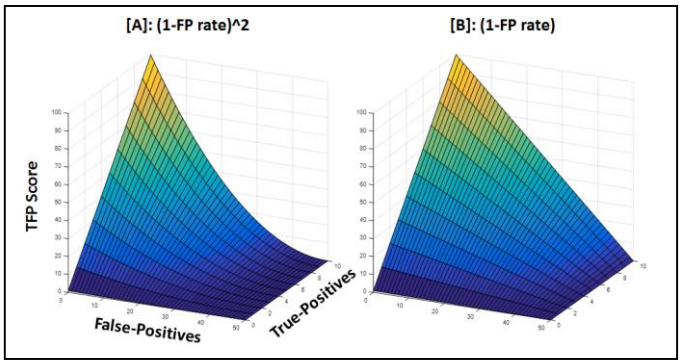

Figure 6. TFP Score graph. A) applies (1-FP rate)^2, as in equation (1), while B) is without the square power. Ranges: TP= 0-10, FP= 0-50, tE=10 and iE=50, TFT=0-100. NB: (1-FP rate)^2 refers to (1- (FP+a)/(iE+a))^2 in eq. 1.

III. Results and Discussion

A. Offline Testing Evaluation

Table I shows classification accuracy for all seven subjects and four different onset tasks. The ***Bold and Italic*** results represent the highest accuracy out of four different onset tasks for each participant. If there is no significant difference between the highest values (as measured by a Wilcoxon test *p*-value), both results are marked as ***Bold and Italic.***

For participants 1, 3 and 7 covert high tone sound-production (C_High) achieved significantly higher accuracy (i.e., average true positive rate when discriminating between idle and task) than the other three tasks (p-value $\ll 0.05$). For participants 2, 3 and 6 the inhibited overt high tone sound-production (IO_High) task achieved the highest accuracy. For participants 4 and 5 there was no significant difference between tasks.

Based on the average values shown at the bottom of Table I, the C_High task led to better results, followed by IO_High, IO_Siren and C_Siren. There was no significant difference between C_High, IO_High and IO_Siren but C_Siren showed significant worse result than other tasks. It is thus advisable to determine the best onset task on an individual basis.

In terms of average performance for each subject, participant 3, 6 and 7 achieved relatively high values. Participant 7 had experience in similar experiments from our previous study in [14], so he/she was expected to achieve high performance. However, participant 3 and 6 were naïve subjects. Also, participants 4 (experienced) and 5 (naïve) showed somewhat low performance results compared to other participants, yet they were experienced users. This suggests that previous experience has no significant effect on performance.

The average true positive rate across all tasks and subjects was 73.76%. However, this value rose to 77.7% if only the best task for each subject was considered.

Table I. Offline testing accuracy from four different sound-production related onset tasks for all subjects

	Accuracy % (Standard Deviation *σ*)				
	C_High	**C_Siren**	**IO_High**	**IO_Siren**	**Average**
P1	***71.98% (±2.66)***	63.56 % (± 5.32)	63.49 % (± 3.77)	***70.20 % (± 3.51)***	67.31 %
P2	73.14 % (± 1.8***)***	68.44 % (± 3.17)	***77.44 % (± 3.51)***	72.41 % (± 3.83)	72.86 %
P3	***87.28 % (± 1.56)***	82.82 % (± 2.07)	***87.20 % (± 2.43)***	78.52 % (± 1.7)	83.96 %
P4	63.42 % (± 2.66)	64.47 % (± 2.05)	62.07 % (± 4.68)	64.89 % (± 3.25)	63.72 %
P5	63.41 % (± 3.75)	64.01 % (± 3.69)	60.43 % (± 4.32)	64.54 % (± 1.87)	63.10 %
P6	73.14 % (± 1.14)	72.69 % (± 1.9)	***85.75 % (± 1.65)***	83.94 % (± 2.64)	78.88 %
P7	***91.84 % (± 1.2)***	80.08 % (± 2.79)	87.53 % (± 1.66)	86.41 % (± 1.32)	86.49 %
Average	74.89 %	70.88 %	74.84 %	74.42 %	73.76 %

B. Simulated-online Testing Evaluation

Figure 7 shows output testing results for participant 6's IO_High onset task, for illustration purposes. It was chosen because the results contain moderate amounts of true-positive and false-positive events, so it allows us to discuss both cases. The horizontal axis represents the time scale in terms of sample windows, one sample representing a 0.5s window.

The vertical axis is binary; a value of 1 indicates a non-idle state, while 0 indicates an idle state. The blue, top line depicts actual onset states as determined from the user's input by pressing space bar after executing a non-idle cognitive task. The green plot (testing output) shows the IC task periods as determined by the LDA classifier.

The red plots (Vote 1 to Vote 6) represent results from an applied voting system, designed to assess sensitivity to false and true events, as follows: Six sequential windows (3s data: 0.5s windows*6) from the testing output were selected and a voting process was applied. Within those 6 sequential windows the machine detected ***N*** onset events. 'Vote ***N'*** denotes the number of onset windows required for the machine to determine that a real onset has occurred. E.g., 'Vote 2' indicates that the machine required 2 (not necessarily consecutive) of the 6 windows to yield 1 as output in order to accept an event as being an onset. This process was continuously done by moving a jumping 0.5s windows (i.e., a sliding window with no overlap) from the beginning to the end of the recorded data. As can be seen from Figure 7, the incidence of false-positives decreases from Vote 1 to Vote 6. However, true-positives also decreased (and in varying degrees, depending on the participant). For this reason, it was necessary to find an optimum voting level to minimise false events while maximising true ones. This was done based on a true-false-positive score (discussed below).

For classification performance assessment, it is difficult to achieve sample by sample labelling in self-paced BCIs as well as in this simulated-online recording protocol. Thus, event by event (i.e., one 0.5s window at a time) labelling was adopted. True-positive and false-positive events were defined as shown in Figure 8. Although participants were instructed to perform a given task for approximately 3s, we included a timing error tolerance region (TETR) to investigate possible timing errors and their effect on system performance. Two different TETR values were investigated in this study, i.e.: the original 3s epoch was padded with the following window lengths on each side: 0.5s (i.e., 0.5s+3s+0.5s = 4s TETR) and 1.5s (6s TETR).

In this experiment, only rising edges from the output graph were only considered as onset. There are three different cases depending on the time location of rising edges. Case 1 indicates the machine-detected rising edge appeared within a TETR and this was treated as true-positive. If there were multiple rising edges in a single TETR (as in case 2), only one true-positive was accepted and others were discarded. Case 3 is an example of a false-positive event. If a rising edge appeared outside the TETR, it was regarded as a false-positive even if remaining machine-detected onset window overlapped with an actual event. If multiple rising edges were detected outside the tolerance region, all of them were considered as false-positives.

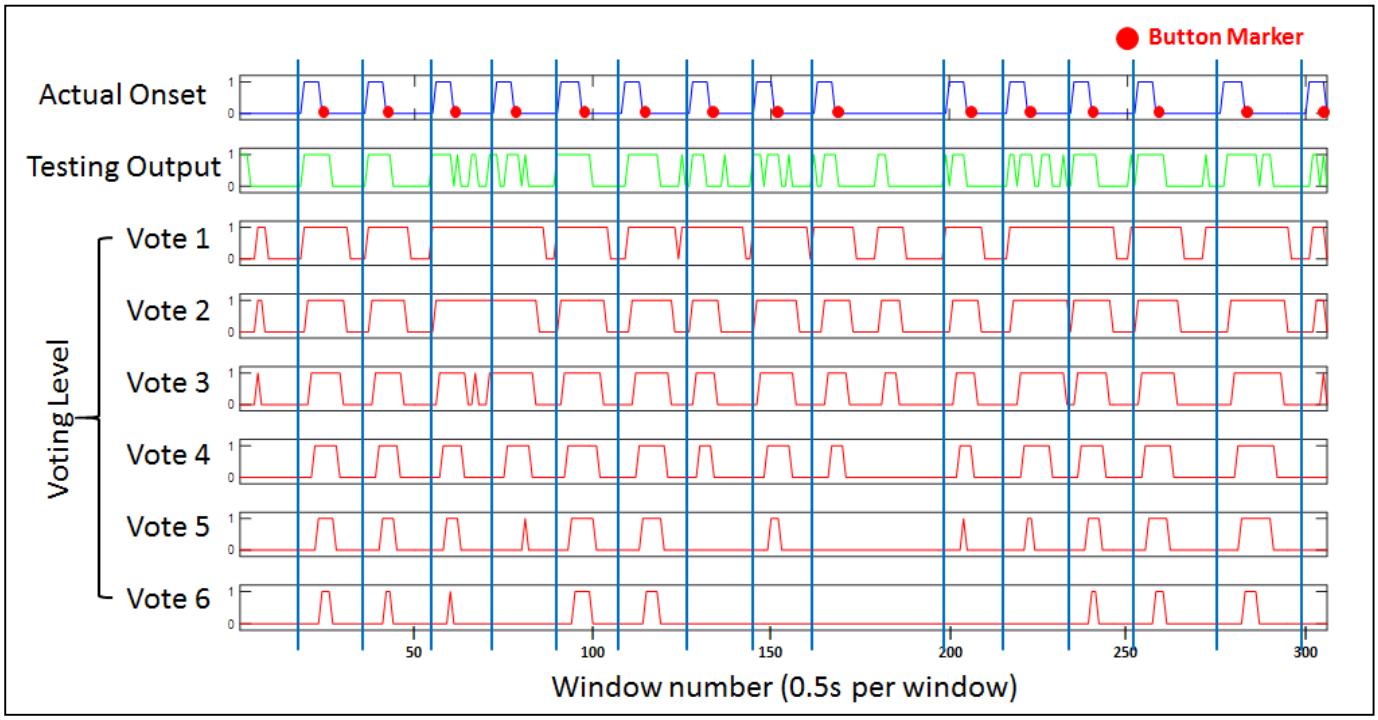

Figure 7. Simulated-online output results for participant 6's inhibited overt high tone onset task. The time scale is shown in terms of sample windows, one sample representing a 0.5s window. 'Button marker' denotes a key press after a 3s task was finished.

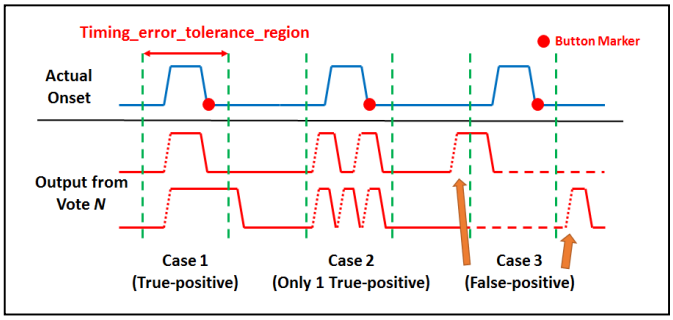

Figure 8. True-positive and False-positive definition in the simulated-online situation.

Table II and Table III show simulated-online testing results for each onset task. The values were calculated based on the true-false-positive score (TFP, described above) and the numbers in a square bracket represent the number of true-positives (TP), false-positives (FP). The total number of actual task onset events (tE) was 15 for all runs. The values shown on the tables for voting level are the ones that gave the highest TFP% score out of six votes.

Two different TETR sizes (4s and 6s TETR) were compared. Larger TETRs increase the chances of detecting true-positive events, while at the same time leading to less frequent false-positives. However, the TFP% score takes into account the total idle period length. Thus, if there was no significant difference in the number of true and false-positives for different TETR values, the smaller TETR, which yields a longer idle period, would give a higher TFP% score**.** The average results showed that 6s TETR (from Table III) has higher score than 4s TETR (from Table II). It leaves us further investigation to find out optimal TETR in usability point of view as a system would give quicker response with smaller TETR. It would be our future study to move online system. In terms of the best voting level sensitivity, results varied widely depending on subject and tasks.

The average TFP score (across participants) for each of the onset tasks were 57.71%, 53.63%, 58.17% and 59.47% (for C_High and C_Siren, IO_High, IO_Siren, respectively) with 4s TETR and 67.79%, 65.10%, 68.49% and 70.13% with 6s TETR. Both results show that IO_Siren task has higher score followed by IO_High, C_High and C_Siren. However, it all vary depends on subjects. When we average the highest TFP scores for each participant, the overall TFP score was 67.12% (TP rate= 72.38%, FP rate=3.78%) with 4s TETR and 76.67% (TP rate= 87.62%, FP rate=4.05%) with 6s TETR.

Table II. Simulated-online performance results. True-false-positive score with optimal voting level. 4s of Timing error tolerance region (TETR).

4s TETR	TFP Score % [TP, FP]				
	C_High	C_Siren	IO_High	IO_Siren	Average
P1	***68.62 % [12, 12]***	42.41 % [7, 8]	45.65 % [8, 12]	68.26 % [12, 12]	56.24 %
P2	***56.60 % [9, 5]***	48.98 % [8, 6]	44.95 % [8, 12]	53.46 % [9, 7]	51.00 %
P3	69.95 % [11, 4]	46.79 % [8, 11]	***78.13 % [12, 2]***	72.55 % [12, 8]	66.86 %
P4	54.06 % [9, 8]	64.08 % [10, 3]	48.90 % [8, 7]	***68.72 % [11, 5]***	58.94 %

P5	49.51 % [8, 6]	49.48 % [8, 6]	***61.73 % [10, 6]***	54.87 % [9, 7]	53.90 %
P6	43.46 % [10, 5]	68.46 % [11, 7]	***74.26 % [12, 7]***	54.56 % [9, 10]	60.19 %
P7	***61.75 % [10, 9]***	55.23 % [9, 9]	53.54 % [9, 6]	43.89 % [7, 8]	53.60 %
Average	57.71 %	53.63 %	58.17 %	59.47 %	57.24 %

Table III. Simulated-online performance results. True-false-positive score with optimal voting level. 6s of Timing error tolerance region (TETR).

6s TETR	**TFP Score % [number of TP, number of FP]**				
	C_High	**C_Siren**	**IO_High**	**IO_Siren**	**Average**
P1	69.57 % [12, 7]	58.00 % [10, 5]	61.87 % [11, 8]	***74.59 % [14, 4]***	66.01 %
P2	***65.42 % [13, 14]***	58.54 % [10, 5]	61.13 % [12, 7]	58.41 % [10, 6]	60.88 %
P3	77.08 % [12, 2]	58.54 % [10, 7]	***88.15 % [14, 3]***	81.87 % [14, 5]	76.41 %
P4	58.85 % [10, 4]	***80.89 % [14, 6]***	54.66 % [9, 6]	79.89 % [12, 0]	68.57 %
P5	***76.70 % [12, 2]***	62.64 % [10, 3]	63.64 % [10, 3]	73.51 % [12, 4]	69.12 %
P6	60.05 % [10, 5]	80.76 % [13, 5]	***86.10 % [14, 4]***	65.79 % [11, 8]	73.18 %
P7	***66.83 % [11, 8]***	56.32 % [9, 8]	63.87 % [11, 13]	56.82 % [9, 6]	60.96 %
Average	67.79 %	65.10 %	68.49 %	70.13 %	67.87 %

C. Comparison with Other Studies

It is very difficult to directly compare our results with other typical motor-imagery onset detection system as there is no common evaluation method. In addition, many studies have shown performance results (such as hit rate) that can only be applied to their own experimental settings (e.g., [4, 5, 40, 41]). Other studies have shown only classification accuracy. In [42] the average TP rate for three subjects for idle vs. motor-imagery was 86.7% and the number of false-positive events was 5.7, but there was no information regarding idle period length, and they also calculated the false-positive rate by treating the number of onset events 'E' as true-negatives, which is a mistake, in our opinion. In [5], motor-imagery versus non-control state achieved classification accuracy around 79.67% on average for three subjects. In [13], six different mental tasks versus idle state achieved around between 55% (Auditory imagery) and 72% (Motor-imagery) offline TP rate on average for 5 subjects. In [43] researchers classified motor-imagery tasks vs. idle state and they used two two-class classifiers for three different classes (left hand and right foot imagery vs. idle). If the feature did not belong to motor-imagery tasks, they assumed it belonged to the idle state. They achieved around 40% true-positive rates in an offline analysis.

Compared to the results from the above studies, our results (i.e., around 76.67% of TFP score, 87.62% TP rate, 4.05% FP rate) look promising. Further, none of the above studies investigated onset timing errors and none attempted to produce a system that would work with a timing error as short as 3s. In addition, our score system based on TFP is more complete and more conservative than previous approaches, making it suitable for future use in asynchronous BCIs.

It is possible that improved results could be obtained by including other wavelet types and orders as well as other feature domains and classifier types. However, we believe that the fact that such simple features (based on the db2 wavelet) and classifier (LDA) yielded encouraging results indicates that the proposed method has potential for further application in BCIs.

IV. CONCLUSIONS

This study presented a methodology to address three current issues in self-paced BCIs: a) determining when an intentional command (IC) is sent by the brain to the machine, b) reliably discriminating between intentional brain activity and brain states that are non-specific or not relevant to the human-machine interaction, and c) the lack of a suitable standard scoring system for performance evaluation in self-paced BCIs.

Averaging all results across all seven participants, the best idle vs. IC offline performance was obtained with the covert high tone (C_High) sound production imagery (74.89% true positive rate, TP rate). 77.7% TP rate was achieved when only the best IC task for each individual participant was used for obtaining average results. These offline results are for a 3s timing window, i.e., a 3s timing uncertainty as to when an actual IC onset occurred. We believe this value is acceptable for most BCI scenarios. For the on-line simulation analysis, IO_siren yielded the best overall results based on the TFP score (68.49%). The average TFP score considering only the best IC task for each participant was 76.67%. The true positive and false positive rates for the latter TFP score were 87.61% and 4.05%, respectively.

While there are no studies against which our results can be directly compared, previous similar IC onset detection studies using motor imagery have yielded best classification (true positive rates) of 72.0% [13] and 79.7% [5], but without taking into account timing errors. In this light, we believe our results, and the proposed methods, may be of use to other self-paced BCI researchers.

REFERENCES

[1] F. Sepulveda, *Brain-actuated Control of Robot Navigation*. INTECH Open Access Publisher, 2011.

[2] Y. Song and F. Sepulveda, "Classifying speech related vs. idle state towards onset detection in brain-computer interfaces overt, inhibited overt, and covert speech sound production vs. idle state," in *Biomedical Circuits and Systems Conference (BioCAS), 2014 IEEE*, 2014, pp. 568-571: IEEE.

[3] C. S. L. Tsui, J. Q. Gan, and S. J. Roberts, "A self-paced brain–computer interface for controlling a robot simulator: an online event

labelling paradigm and an extended Kalman filter based algorithm for online training," *Medical & biological engineering & computing*, vol. 47, no. 3, pp. 257-265, 2009.

[4] R. Leeb, D. Friedman, G. R. Müller-Putz, R. Scherer, M. Slater, and G. Pfurtscheller, "Self-paced (asynchronous) BCI control of a wheelchair in virtual environments: a case study with a tetraplegic," *Computational intelligence and neuroscience*, vol. 2007, 2007.

[5] R. Scherer, F. Lee, A. Schlogl, R. Leeb, H. Bischof, and G. Pfurtscheller, "Toward self-paced brain–computer communication: navigation through virtual worlds," *Biomedical Engineering, IEEE Transactions on*, vol. 55, no. 2, pp. 675-682, 2008.

[6] M. Fatourechi, R. Ward, and G. Birch, "A self-paced brain–computer interface system with a low false positive rate," *Journal of neural engineering*, vol. 5, no. 1, p. 9, 2008.

[7] K. Brigham and B. Kumar, "Imagined speech classification with EEG signals for silent communication: a preliminary investigation into synthetic telepathy," in *Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference on*, 2010, pp. 1-4: IEEE.

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

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

[10] L. Wang, X. Zhang, X. Zhong, and Y. Zhang, "Analysis and classification of speech imagery eeg for bci," *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 901-908, 2013.

[11] C. Herff *et al*., "Brain-to-text: decoding spoken phrases from phone representations in the brain," *Frontiers in Neuroscience*, vol. 9, Jun 12 2015, Art. no. 217.

[12] J. Roland, P. Brunner, J. Johnston, G. Schalk, and E. C. Leuthardt, "Passive real-time identification of speech and motor cortex during an awake craniotomy," *Epilepsy & Behavior*, vol. 18, no. 1, pp. 123-128, 2010.

[13] M. Dyson, F. Sepulveda, J. Q. Gan, and S. J. Roberts, "Sequential classification of mental tasks vs. idle state for EEG based BCIs," in *Neural Engineering, 2009. NER'09. 4th International IEEE/EMBS Conference on*, 2009, pp. 351-354: IEEE.

[14] Y. J. Song and F. Sepulveda, "Classifying siren-sound mental rehearsal and covert production vs. idle state towards onset detection in brain-computer interfaces," in *Brain-Computer Interface (BCI), 2015 3rd International Winter Conference on*, 2015, pp. 1-4: IEEE.

[15] S. Martin *et al.*, "Decoding spectrotemporal features of overt and covert speech from the human cortex," *Frontiers in neuroengineering*, vol. 7, 2014.

[16] J. O. Pickles, *An introduction to the physiology of hearing*. BRILL, 2012.

[17] K. R. Sitek, D. H. Mathalon, B. J. Roach, J. F. Houde, C. A. Niziolek, and J. M. Ford, "Auditory cortex processes variation in our own speech," 2013.

[18] K. B. Ng, A. P. Bradley, and R. Cunnington, "Effect of competing stimuli on SSVEP-based BCI," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, 2011, pp. 6307-6310: IEEE.

[19] C. Cinel, R. Poli, and L. Citi, "Possible sources of perceptual errors in P300-based speller paradigm," 2004.

[20] X. Pei, E. C. Leuthardt, C. M. Gaona, P. Brunner, J. R. Wolpaw, and G. Schalk, "Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition," *Neuroimage*, vol. 54, no. 4, pp. 2960-2972, 2011.

[21] V. L. Towle *et al.*, "ECoG gamma activity during a language task: differentiating expressive and receptive speech areas," *Brain*, vol. 131, no. 8, pp. 2013-2027, 2008.

[22] N. Crone *et al.*, "Electrocorticographic gamma activity during word production in spoken and sign language," *Neurology*, vol. 57, no. 11, pp. 2045-2053, 2001.

[23] M. Kirkove, C. François, and J. Verly, "Comparative evaluation of existing and new methods for correcting ocular artifacts in electroencephalographic recordings," *Signal Processing*, vol. 98, pp. 102-120, 2014.

[24] V. Krishnaveni, S. Jayaraman, S. Aravind, V. Hariharasudhan, and K. Ramadoss, "Automatic identification and Removal of ocular artifacts from EEG using Wavelet transform," *Measurement science review*, vol. 6, no. 4, pp. 45-57, 2006.

[25] B. Perseh and A. R. Sharafat, "An efficient P300-based bci using wavelet features and IBPSO-based channel selection," *Journal of medical signals and sensors*, vol. 2, no. 3, p. 128, 2012.

[26] I. Güler and E. D. Übeyli, "Adaptive neuro-fuzzy inference system for classification of EEG signals using wavelet coefficients," *Journal of neuroscience methods*, vol. 148, no. 2, pp. 113-121, 2005.

[27] A. Subasi, "EEG signal classification using wavelet feature extraction and a mixture of expert model," *Expert Systems with Applications*, vol. 32, no. 4, pp. 1084-1093, 2007.

[28] T. Gandhi, B. K. Panigrahi, and S. Anand, "A comparative study of wavelet families for EEG signal classification," *Neurocomputing*, vol. 74, no. 17, pp. 3051-3057, 2011.

[29] J. C. Bezdek and N. R. Pal, "Some new indexes of cluster validity," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 28, no. 3, pp. 301-315, 1998.

[30] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 2, pp. 224-227, 1979.

[31] F. Lotte, M. Congedo, A. Lécuyer, and F. Lamarche, "A review of classification algorithms for EEG-based brain–computer interfaces," *Journal of neural engineering*, vol. 4, 2007.

[32] M. Fatourechi, A. Bashashati, R. K. Ward, and G. E. Birch, "EMG and EOG artifacts in brain computer interface systems: A survey," *Clinical neurophysiology*, vol. 118, no. 3, pp. 480-494, 2007.

[33] A. Vergult *et al*., "Improving the interpretation of ictal scalp EEG: BSS–CCA algorithm for muscle artifact removal," *Epilepsia*, vol. 48, no. 5, pp. 950-958, 2007.

[34] W. De Clercq, A. Vergult, B. Vanrumste, W. Van Paesschen, and S. Van Huffel, "Canonical correlation analysis applied to remove muscle artifacts from the electroencephalogram," *Biomedical Engineering, IEEE Transactions on*, vol. 53, no. 12, pp. 2583-2587, 2006.

[35] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321-377, 1936.

[36] O. Friman, M. Borga, P. Lundberg, and H. Knutsson, "Exploratory fMRI analysis by autocorrelation maximization," *NeuroImage*, vol. 16, no. 2, pp. 454-464, 2002.

[37] D. Safieddine *et al.*, "Removal of muscle artifact from EEG data: comparison between stochastic (ICA and CCA) and deterministic (EMD and wavelet-based) approaches," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, pp. 1-15, 2012.

[38] J. Gao, C. Zheng, and P. Wang, "Online removal of muscle artifact from electroencephalogram signals based on canonical correlation analysis," *Clinical EEG and neuroscience*, vol. 41, no. 1, pp. 53-59, 2010.

[39] G. Townsend, B. Graimann, and G. Pfurtscheller, "Continuous EEG classification during motor imagery-simulation of an asynchronous BCI," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 12, no. 2, pp. 258-265, 2004.

[40] B. Xia, D. An, C. Chen, H. Xie, and J. Li, "A mental switch-based asynchronous brain-computer interface for 2D cursor control," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, 2013, pp. 3101-3104: IEEE.

[41] Y. Chae, S. Jo, and J. Jeong, "Brain-actuated humanoid robot navigation control using asynchronous brain-computer interface," in *Neural Engineering (NER), 2011 5th International IEEE/EMBS Conference on*, 2011, pp. 519-524: IEEE.

[42] C. Tsui, A. Vučković, R. Palaniappan, F. Sepulveda, and J. Gan, *Narrow band spectral analysis for movement onset detection in asynchronous BCI*. na, 2006.

[43] D. Zhang, Y. Wang, X. Gao, B. Hong, and S. Gao, "An algorithm for idle-state detection in motor-imagery-based brain-computer interface," *Computational intelligence and neuroscience*, vol. 2007, pp. 5-5, 2007.