Toward an Imagined Speech-Based Brain Computer Interface Using EEG Signals

Mashael M. Alsaleh

Department of Computer Science

The University of Sheffield

This thesis is submitted for the degree of

Doctor of Philosophy

Speech and Hearing Research Group

July 2019

I would like to dedicate this thesis to

My loving parents, thank you for everything.

Ever loving memory of my sister Norah, until we meet again.

Declaration

I hereby declare that I am the sole author of this thesis. The contents of this thesis are my original work and have not been submitted for any other degree or any other university. Some parts of the work presented in Chapters 3, 4, 5, and 6 have been published in conference proceedings, and a journal as given below:

- AlSaleh, M. M., Arvaneh, M., Christensen, H., and Moore, R. K. (2016, December). Brain-computer interface technology for speech recognition: a review. In 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA) (pp. 1-5). IEEE.
- AlSaleh, M., Moore, R., Christensen, H., and Arvaneh, M. (2018, July). Discriminating Between Imagined Speech and Non-Speech Tasks Using EEG. In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 1952-1955). IEEE.
- Alsaleh, M., Moore, R., Christensen, H., and Arvaneh, M. (2018, October). Examining Temporal Variations in Recognizing Unspoken Words Using EEG Signals. In 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 976-981). IEEE.
- Alsaleh, M., Moore, R., Christensen, H., and Arvaneh, M. (2019). EEG-based Recognition of Unspoken Speech using Dynamic Time Warping. planned journal publication.

Mashael M. Alsaleh

July 2019

Acknowledgements

In the name of ALLAH, the Most Gracious and the Most Merciful. Thanks to ALLAH who is the source of all the knowledge in this world, for the strengths and guidance in completing this thesis. I thank those who rendered their direct or indirect support during the period of my PhD work.

First of all, I would like express my utmost gratitude to my supervisors, Professor Roger K. Moore and Dr. Mahnaz Arvaneh, for their mentorship, guidance and support throughout my time studying in Sheffield. Their constructive comments and suggestions, throughout the experimental and thesis work, have contributed to the success of this research. It has been an honour and privilege to work with them. Special thanks to my panel members, Dr. Heidi Christensen and Dr. Anthony Simons for their useful suggestions and for being so kind to me.

I also like to thank my examiners, Prof. Tomas Ward and Dr. Yoshi Gotoh for the valuable comments and an intellectually stimulating and enjoyable viva.

I am hugely grateful to my parents for their support, prayers, love and care throughout my life; they have played a vital role in helping me to reach this milestone. I owe special thanks to my sister Nouf, my brothers Abdullah and Saad for their continuous love and prayers. My beautiful niece Norah, you came into our family bringing much joy and love.

My friends who never let me feel that I am away from my homeland: Rabab, Sadeen, Lubna, Hanan, Maryam, and Najwa. The members of the Speech and Hearing lab and Physiological Signals and Systems lab at University of Sheffield have been, and continue to be, a source of knowledge, friendship and humour.

Last but not the least, I thankfully acknowledge the financial funding for this work from King Saud University in Saudi Arabia, who gave me a scholarship to pursue my PhD studies.

Abstract

Individuals with physical disabilities face difficulties in communication. A number of neuromuscular impairments could limit people from using available communication aids, because such aids require some degree of muscle movement. This makes brain–computer interfaces (BCIs) a potentially promising alternative communication technology for these people. Electroencephalographic (EEG) signals are commonly used in BCI systems to capture non-invasively the neural representations of intended, internal and imagined activities that are not physically or verbally evident. Examples include motor and speech imagery activities.

Since 2006, researchers have become increasingly interested in classifying different types of imagined speech from EEG signals. However, the field still has a limited understanding of several issues, including experiment design, stimulus type, training, calibration and the examined features. The main aim of the research in this thesis is to advance automatic recognition of imagined speech using EEG signals by addressing a variety of issues that have not been solved in previous studies. These include (1) improving the discrimination between imagined speech versus non-speech tasks, (2) examining temporal parameters to optimise the recognition of imagined words and (3) providing a new feature extraction framework for improving EEG-based imagined speech recognition by considering temporal information after reducing within-session temporal non-stationarities.

For the discrimination of speech versus non-speech, EEG data was collected during the imagination of randomly presented and semantically varying words. The non-speech tasks involved attention to visual stimuli and resting. Time-domain and spatio-spectral features were examined in different time intervals. Above-chance-level classification accuracies were achieved for each word and for groups of words compared to the non-speech tasks.

To classify imagined words, EEG data related to the imagination of five words was collected. In addition to words classification, the impacts of experimental parameters on classification accuracy were examined. The optimization of these parameters is important to improve the rate and speed of recognizing unspoken speech in on-line applications. These parameters included using different training sizes, classification algorithms, feature extraction in different time intervals and the use of imagination time length as classification feature. Our extensive results showed that Random Forest classifier with features extracted using Discrete Wavelet Transform from 4 seconds fixed

time frame EEG yielded that highest average classification of 87.93% in classification of five imagined words.

To minimise within class temporal variations, a novel feature extraction framework based on dynamic time warping (DTW) was developed. Using linear discriminant analysis as the classifier, the proposed framework yielded an average 72.02% accuracy in the classification of imagined speech versus silence and 52.5% accuracy in the classification of five words. These results significantly outperformed a baseline configuration of state-of-the art time-domain features.

Table of contents

List of figures

List of tables

Abbreviation

ANN Artificial Neural Network.

AR Autoregressive.

BCI Brain Computer Interface.

CC Cross-Correlation.

CNS Central Nervous System.

CSP Common Spatial Patterns.

DDTW Derivative Dynamic Time Warping.

DTW Dynamic Time Warping.

DWT Discrete Wavelet Transform.

ECoG Electrocorticography.

EEG Electroencephalogram.

EMG Electromyogram.

EOG Electrooculogram.

ERD/ERS Event-Related Desynchronization/Synchronization.

ERPs Event Related Potentials.

ErrPs Error-Related Potentials.

FBCSP Filter-bank common spatial patterns.

FD Frequency domain.

fMRI Functional magnetic resonance imaging.

fNIRS Functional Near Infrared Spectroscopy.

ICA Independent Component Analysis.

KNN k-nearest-neighbours.

LDA Linear discriminant analysis.

MaxCC Maximum Cross-Correlation.

MEG Magnetoencephalography.

NB naive Bayes.

RF Random Forest.

RMS Root Mean Square.

RWE Relative Wavelet Energy.

SCPs Slow Cortical Potentials.

SD Standard Deviation.

SSVEP Steady State Evoked Potential.

SVM Support Vector Machine.

TD Time domain.

TFD Time-Frequency domain.

Chapter 1

Introduction

Most communication aids require some degree of muscle control, which makes such aids unsuitable for people with severe motor disabilities, such as those with locked-in syndrome. The use of a brain-computer interface (BCI) might be the only option for such users, since BCI does not require any muscular activity. For these patients, the inability to communicate verbally has wide-ranging impacts and can result in significantly reduced social interaction and possible isolation. At the same time, caregivers have much more difficulty in determining their patients' needs. Such concerns have been key in the development of BCI communication technologies (Brumberg et al., 2011; Oken et al., 2014). In addition to improve communication for individuals with physical disabilities, EEG could be potentially used to augment conventional communication for healthy users. This would be preformed by reading users thoughts by typing straight from brain which is five times faster than typing on phone. This idea has been proposed and discussed by Facebook [1].

BCI systems work based on measuring specific features of brain activities related to the user's intention and then translating those features into device-control signals. Several alternative cognitive control instructions may be used to capture the user's intention through the use of perceptual or imagination tasks. Among them, speech imagination is the most intuitive and is closer to the natural communication pathway in

[1]BBC News: https://www.bbc.co.uk/news/technology-39648788.

comparison to other types of BCI, such as motor imagery (Brunner et al., 2008), steady state visual evoked potential (SSVEP) (Chen et al., 2015) and P300 (Van Gerven et al., 2009). Speech imagination does not require any external stimuli and, if classified accurately, can also support a larger number of classes than the better known motor imagination technique. For example, motor imagination is currently known to be relatively accurate using a maximum of four classes (Brunner et al., 2008). However, these four classes of imaginations are not directly related to verbal communication.

Torres-García et al. (2016) defined 'imagined speech' as the internal pronunciation of vocabulary with no muscle involvement or sound production. Morin and Michaud (2007) listed other closely related terms for imagined speech, including "*self-talk, sub-vocal/covert speech, internal dialogue/ monologue, sub-vocalization, utterance, self-verbalization, and self-statement*". Other studies used the term 'unspoken speech'. Figure 1.1 shows a comparison of different speech modalities and how unspoken speech differs from these modalities.

Beyond communication, imagined speech also occurs in everyday life. Researchers have described imagined speech as the foundation for reviewing short-term memory (Baddeley et al., 1975) and providing a phonological effect during reading and writing (Oppenheim and Dell, 2008). Dolcos and Albarracin (2014) suggested that imagined speech may be employed to represent, maintain, and arrange task-related data and conscious thought processes.

A much-debated question is how imagined speech happens. Some studies in the literature have proposed that imagined speech is generated in a similar manner to overt speech but without articulator movement that leads to audio (Oppenheim and Dell, 2010). Martínez-Manrique and Vicente (2015) argued that imagined speech is not an independent cognitive function but instead inherits functions from overt speech.

Research on imagined speech using non-invasive BCI started in 1997 (Suppes et al., 1997). In that study, the researchers combined electroencephalogram (EEG) and magnetoencephalography (MEG) in order to classify imagined words. Since that time until the research for this thesis was begun in 2015, only a few other studies have been

conducted. As a result, several research gaps and questions still need to be answered before this technology may be extended to its intended users.

The remainder of this chapter is organized as follows. The research gaps and challenges are presented in section 1.1, while the research aims and objectives of this thesis are described in section 1.2. Finally, section 1.3 lists the chapters that comprise the remainder of this thesis.

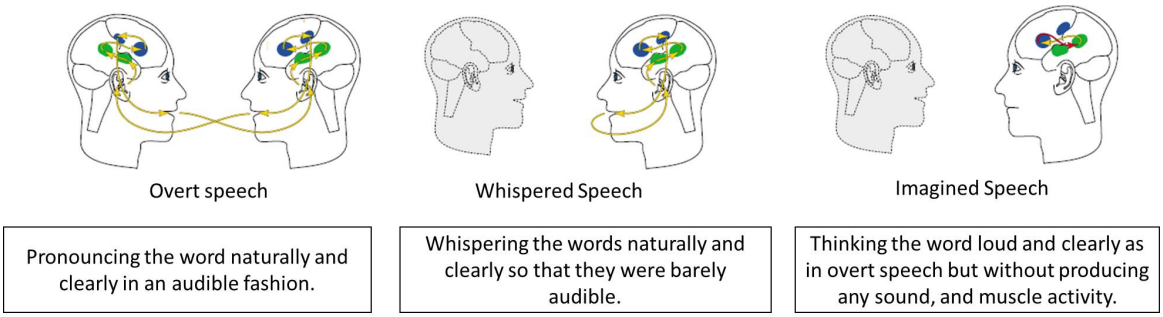

Fig. 1.1 The differences between three different speech modalities; adapted from (Hesslow, 2002)

1.1 Research challenges in speech recognition using EEG signals

The challenges encountered during the pursuit of this PhD thesis may be divided into two parts. The first is the challenge of providing robust high-classification accuracy, as such accuracy is crucial for providing a technology that will help patients. This challenge arises because of the nature of EEG signals, which have poor spatial resolution and are easily contaminated by noise. The second challenge is that the types of EEG patterns investigated in this thesis are a relatively new topic. Speech imagination has not been studied much in the literature compared to topics such as motor imagination. The following section describes these two challenges with further examples.

1.1.1 The nature of EEG signals

The spatial resolution of EEG signals is generally poor (Nunez et al., 1997). A number of neural activities known as central nervous system (CNS) noise, that are unrelated but occur at the same time, usually affect the information of interest due to overlapped features (Wolpaw et al., 2002). For instance, event-related desynchronization/synchronization (ERD/ERS) patterns are generally detected in the μ-rhythm (i.e. 8–13 Hz) in the context of motor-imagery-based BCIs (Pfurtscheller and Da Silva, 1999). The frequency range of the μ-rhythm is identical to that of the α-rhythm, however, which is indicative of visual and mental activity. The occipital brain lobe is the origin of the α-rhythm, but the α-rhythm mixes with the μ-rhythm that emerges from the motor cortex because of volume conduction (Nunez et al., 1997).

Non-stationarity is another issue to affect EEG signal classification. Non-stationarity, which refers to the fact that EEG properties vary between intra- and inter-sessions, can lower BCI performance. The majority of machine-learning algorithms are based on the assumption of data stationarity (Shenoy et al., 2006). This situation means that, if the classifier is matched to the training data of the subject, the classifier may be inappropriate for new session data on a different day or even for new trials in the same session (Kawanabe et al., 2010). The changes in EEG signal dynamics are related to several causes (Shenoy et al., 2006). First, the physical qualities of EEG electrodes decrease over time. For instance, when the conductive gel dries up, the electrode impedance can become altered, or the reuse of the EEG cap in a new session can cause the electrodes to shift position. Secondly, variations in neurophysiological conditions, such as wakefulness, can occur over time. Thirdly, significant variations arise from psychological variables such as motivation, attention, and task participation. Fourthly, the signal properties may be altered by artefacts that result from body motions or muscle activity, such as swallowing or blinking. Finally, non-stationarity can arise from neuro-feedback as well; the attained neuro-feedback suggests that users attempt to improve outcomes by modifying their brain patterns.

It is also important to note that the imagination of different words involve activation of different areas of the brain such as auditory cortex, Wernicke's area, Broca's area, and the angular gyrus (see Section 2.1 for more information about speech production). These activations overlapping a lot due to volume conduction and poor spatial resolution of EEG. Thus, there is a need for advanced signal processing and machine learning algorithms to accurately classify such mixed EEG patterns.

The ongoing usage of BCI systems can clearly be obstructed by the shortcomings outlined above. In real-life settings, BCIs must demonstrate precision and robustness in application. Patients require a system that will be capable of adjusting to new situations and that will never malfunction.

1.1.2 Speech imagery research

As was described earlier in the introduction to this chapter, the examination of imagined speech using non-invasive technology has received scant attention in the research literature. A few recent studies were conducted late in 2017, but building solid conclusions and finding a baseline to build upon are still difficult tasks. The following are summary of the limitations and gaps in the field of speech imagination. More details and discussions about the previous studies are in Chapter 3.

First, there is a lack of publicly available EEG datasets of speech imaginations of English words. Second, researchers have argued about the optimal design for collecting EEG patterns related to speech. The first study to use EEG signals only was by Wester (2006). The data recording in that study was based on blocks, where each word was repeated in several trials consequently. Porbadnigk et al. (2009), however, have argued that this approach adds time correlation to the EEG patterns, which makes the recognition connected to time instead of speech . In further studies, some authors have followed the block-recording method, while others have performed the recording using different approaches. These approaches are discussed further in Chapter 3. Other negotiable design perspectives involve which types of stimuli are more recognizable

(vowels, syllables, or words), the length of imagination tasks, and how best to separate the imagined patterns.

Third, the limited data and the few studies that have been conducted on the subject have led to few features and perspectives being examined in EEG for imagined speech. Most of the features that have been examined were focused on time-frequency features. Few researchers have paid attention to time-domain features. Most of the previous studies failed in considering temporal variations between EEG trials. Such variations can happen as a result of several factors including, the differences in the imagination start time, and the speed of imagination. Considering these variations can improve the discrimination between imagined speech and non-speech tasks and between different imagined words.

Finally, researchers have argued about the imagined speech process, i.e. which areas of the brain contribute to imagined speech and which brain rhythms are important. Unlike motor imagery, these issues have not been well-examined and studied.

Based on the research gaps outlined above two research questions can be highlighted to be addressed by the research in this thesis. **The first question was** *What are the key parameters to optimise in order to enhance the classification of imagined speech versus non-speech and the classification of imagined words.* **The second question was** *does minimising the temporal variations between the imagined speech trials enhance the classification of imagined speech versus silence and between different imagined words.*

1.2 Research aim and objectives

The main aim of this research is to advance EEG-based BCIs for the recognition of unspoken speech. Considering this aim and the research questions outlined above, the following objectives were set as intermediate steps towards that goal.

The first objective was to discriminate between imagined speech and two types of non-speech tasks related to either a visual stimulus or relaxation. This stage involves the collection of the EEG data of imagined speech for a variety of semantically different

words. The classification of these words was then examined against non-speech tasks where time-domain and spatio-spectral features are examined in different time intervals and using different classifiers.

The second objective involved discriminating between five different imagined words. This stage involved (a) optimizing parameters related to experimental paradigms, such as type/order of stimulus and training size, and (b) optimizing the computational model for word recognition, such as using different time intervals for feature extraction and by examining the use of word length as classification feature.

For the third objective, temporal variations were investigated over time in the recognition of words. Most of the features that have been used in previous studies are time-frequency features. Numerous techniques are available for managing temporal information. In this thesis, dynamic time warping (DTW) is used to examine temporal information in EEG patterns related to imagined speech. DTW measures similarities between two signals by compressing/expanding the signals and by looking for the best non-linear temporal alignment. DTW was originally introduced in the field of audio speech processing. Only a few studies have applied DTW for alignment of brain patterns, such as EEG (Chaovalitwongse and Pardalos, 2008), and Electrocorticography (ECoG) (Martin et al., 2016). This use of DTW for imagined speech recognition involves the following:

- Propose a novel framework for feature extraction using DTW. To the best of the researcher's knowledge, this thesis presents the first use of DTW in the context of EEG for imagined speech.
- Evaluate the proposed framework and compare it with state-of-the-art algorithms.

1.3 Thesis structure

The remainder of this thesis is structured as follows:

- **Chapter 2: Brain Computer Interface for Communication.** This chapter provides extensive background information about BCI technology in general,

with a particular concentration on communication applications. The chapter includes a short review of brain areas and their relations to speech production and discusses the relation between overt and imagined speech.

- **Chapter 3: Brain Computer Interface for Unspoken Speech Recognition.** This chapter discusses studies that have been conducted in the context of imagined speech using either invasive or non-invasive BCI technologies. The chapter's most detailed investigation is on EEG for imagined speech, as this is the technology of interest in this thesis. Finally, the chapter provides a discussion of current state-of-the-art studies and includes possible areas for improvement.

- **Chapter 4: Discriminating between Imagined Speech and Non-speech.** This chapter presents the experiment that was conducted to achieve the thesis's first objective. The chapter describes the motivation for this experiment and compares this work with studies from the literature. The chapter then describes the data-collection process and the proposed feature-extraction and classification algorithms. Finally, the chapter discusses the results and proposes enhancements to the experimental design.

- **Chapter 5: Examining Temporal Issues Related to Unspoken Speech Recognition.** This chapter is related to the second research objective. First, the chapter presents first classification between five imagined words. Second, questions related to temporal issues in the experimental design are answered. The chapter starts with a discussion of the variations in the experimental design from previous studies. It then lists the experiment's motivations and research questions before providing the data-collection process and the experimental design. Finally, the chapter provides answers to the questions introduced earlier in the chapter.

- **Chapter 6: Dynamic Time Warping in the Recognition of Unspoken Speech.** This chapter proposes and evaluates DTW-based feature extraction framework (the thesis's third objective). The chapter first discusses the importance of temporal information and time-domain features before introducing DTW

and how it works. The chapter then reviews previous BCI studies that have used DTW. It then presents the proposed DTW feature-extraction framework. The chapter then presents the methodology for framework evaluation before stating and discussing the results.

- **Chapter 7: Conclusion.** This chapter summarizes the main contributions of this thesis and considers the limitations of the study. The chapter also presents possible future work.

Chapter 2

Brain Computer Interface for Communication

BCIs have been the subject of cognitive neuroscience research since the 1980s. The first BCI international meeting defined the term as "*A brain-computer interface is a communication system that does not depend on the brain's normal output pathways of peripheral nerves and muscles*" Wolpaw et al. (2000). BCI is used to assist people who are entirely paralysed and so have limited use of their brain activity signals, meaning they have no control over their muscles. BCI makes use of brain patterns linked with conscious or unconscious brain activities, thus offering a way for patients to communicate when their control of their motor system and muscles is too poor to allow them to communicate in the usual manner.

This chapter presents a comprehensive overview of brain-computer interfaces, with a focus on communication applications. Though intended for non-specialists, the chapter contains the technical and background details for subjects discussed in the remainder of the thesis. After providing an overview of the brain areas related to language, core components of the BCI systems will be explained. The chapter then describes possible techniques for measuring brain activities. The discussion then moves to EEG, which provides the signals used for BCI that will be discussed in the remainder of the thesis. Different types of neurophysiological signals in EEG are presented using examples of

these signals' applications as assistive communication tools. Finally, various possible applications of BCI are described.

2.1 Brain and language

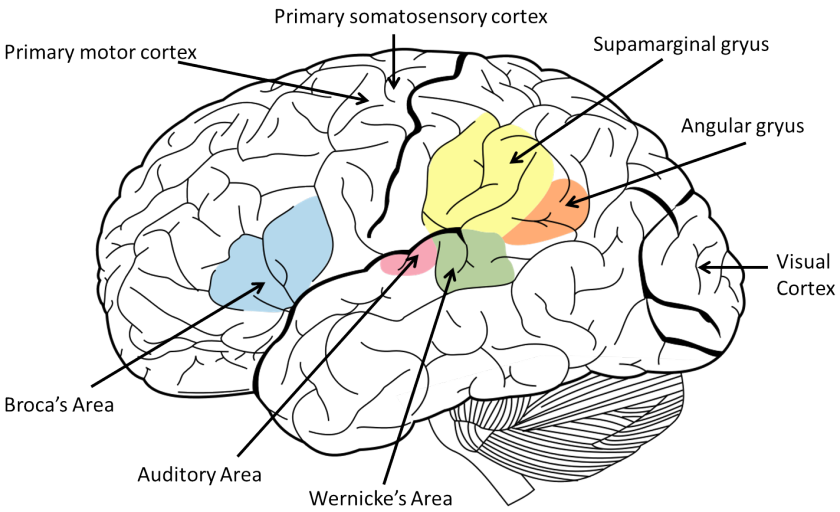

Fig. 2.1 Area in brain that are related to language and speech (Wiki Commons-released to the public domain)

Brain areas and their respective functions do not have a simple one-to-one link. One famous misunderstanding is that a single brain area is responsible for processing vision, another for smelling, and another for language; in reality, brain functions are much more complicated (Vaadia and Birbaumer, 2009). Speech and language functions are spread across many areas of the brain, with crucial parts found in every section of the brain. The *perisylvian* language zone of the dominant language hemisphere, which is predominantly found in the left hemisphere of the brain, is the area most often linked with language. The zone has been shown to control speech for 96% of right-handed people, and 70% of left-handed people (Gick et al., 2012). The term *perisylvian* describes areas surrounding the Sylvian fissure, including the auditory cortex, Wernicke's area, Broca's area, and the angular gyrus, as depicted in Figure 2.1.

- **Broca's area**:

  Paul Broca first found the area now named for him in 1861 during autopsy work Broca (1861). Broca saw that this area showed signs of injury for people who were unable to effectively articulate words. In certain instances, they were able to speak only a few words in total. Today, this area of the brain is widely believed to account for the articulation of words.

- **Wernicke's area**:

  The Wernicke's area is found in the left side of the brain. Wernicke (1974) discovered the area after noting that a lesion in the area brought on speech without language. As a result, people would be able to speak fluently without making any sense; they could form words without any meaning into sentences that sounded reasonable.

- **Angular gyrus and supramarginal gyrus**:

  They can be found in bilaterally in the parietal lobe, close to the superior edge of the temporal lobe. They are responsible for processing high level information of speech such as phonological processing and emotional responses. Angular gyrus is responsible for a person's ability to read and write as well as being involved in perception for the multi-modal integration of speech information (Gick et al., 2012).

- **The primary somatosensory cortex**:

  It can be found in the partial lobe, and it is responsible for processing of tactile information during speech perception and feedback system in speech production (Bertelson et al., 2003).

- **The visual cortex**:

  It is located in the occipital lobe at the posterior of the brain, and is responsible for processing visual speech information, and responding to visual speech while listening to speech in noisy environment (Schepers et al., 2014).

- **Primary motor cortex**:

  As shown in Figure 2.2, primary motor cortex controls movement in most of the human body, including the vocal speech tract. As the figure shows, certain sections of the motor cortex manage specific body parts; the size of the body in the figure is not actual size but is related to the brain portion in control of that specific body part. Broca's area produces a speech plan, as released to the primary motor cortex. The plan is then sent to the lower brain areas, where it is spread to the various body parts, which, in turn, will be moved (Gick et al., 2012).

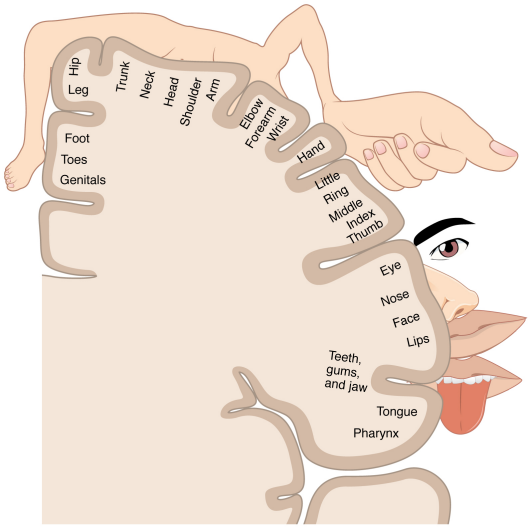

Fig. 2.2 Illustration of the layout of the cortical map in the primary motor cortex (Wiki Commons- released to the public domain)

2.1.1 Process of speech production in the human brain

The way in which the brain creates speech is an area of keen research interest. The *Wernicke-Geschwind* model, which is an important aspect of this process, helps to describe a commonly held theory regarding the production of speech once a person hears a word (Geschwind, 1979). The model describes the processes involved after someone hears a word and then wishes to say that word. The word is first processed in

the primary auditory area, where semantics are taken and added within the Wernicke's area. As the signal progresses across the *arcuate fasciculus* (the link joining the Broca's area with the Wernicke's area), a plan for the motor cortex is created in the Broca's area. This plan is then used in the motor cortex, which makes use of the vocal tract in an appropriate way.

2.1.2 The similarity between voiced and imagined speech production

There is currently no agreement on the exact nature of the association between voiced speech and imagined speech (Brocklehurst and Corley, 2011; Oppenheim and Dell, 2008, 2010). This section first presents previous examinations of the association between voiced and imagined speech before reviewing the differences between these two forms of speech communication.

Oppenheim and Dell (2010) proposed that imagined speech is an abbreviated form of voiced speech. In both forms of speech, identical phases of speech production occur. Subjective descriptions of imagined speech reveal that it is similar to voiced speech in rhythm, tempo, and pitch (MacKay, 2014). According to the motor simulation hypothesis, voiced and imagined speech involve similar linguistic processes and physiological correlates (Perrone-Bertolotti et al., 2014). Although imagined speech displays somewhat weakened features (Alderson-Day and Fernyhough, 2015).

Phonemic similarity, has been noted in relatively similar magnitudes in both voiced and imagined speech generation (Brocklehurst and Corley, 2011). Other reports have indicated that imagined speech is stated at the sub-phonemic level with speech generation processes that resemble those of voiced speech (Corley et al., 2011). This similarity implies that imagined speech retains the same degree of featural richness as voiced speech and that phonological information are included in imagined speech.

Marvel and Desmond (2012) reported a high similarity between voiced and imagined speech neurobiology. They found that neural activations typically took place in left-brain hemispheric language regions and were generally correlated with both forms

of speech communication (Basho et al., 2007; McGuire et al., 1996; Palmer et al., 2001). The activation of Broca's area, which occurs in imagined speech, reveals that this typical language region of the brain is involved with imagined speech generation. This theory has been validated by findings from functional imaging studies on silent articulation (Paulesu et al., 1993). While Oppenheim and Dell (2008) proposed a major overlap between voiced and imagined speech, they also suggested that imagined speech is relatively minimal at the level of features. Therefore, it is both abstract and underspecified. Researchers have also proposed that imagined speech, which is frequently diminutive at the superficial level, lacks phonological (Oppenheim and Dell, 2008) or phonetic (Wheeldon and Levelt, 1995) detail.

The counterview that imagined speech is inherently similar to voiced speech is known as the abstraction hypothesis. This theory proposes that imagined speech is generated as a result of the activation of the representations of abstract linguistic (Indefrey and Levelt, 2004). According to this theory, imagined speech starts prior to the speaker retrieving articulatory data, and such speech should therefore not need any motor activations. Several justifications support this abstraction view, as explained by Oppenheim and Dell (2010). Therefore, it lacks the same articulatory characteristics as voiced speech. Second, in language-related brain regions, the attenuated activity that occurs while generating imagined speech indicates generation mechanisms that are less complete than in voiced speech. A third argument suggests that articulatory abilities are not required for imagined speech. According to this view, articulatory suppression does not necessarily lead to the suppression of imagined speech. Furthermore, imagined speech need not lead to voiced speech articulation. Theoretically, if voiced and imagined speech involve identical planning mechanisms, then the practice of one form of speech could enhance performance in the other.

As mentioned earlier, previous studies have noted neuro-anatomical overlap between the regions of the brain correlated with voiced and imagined speech. That being said, there are important variations in brain activity between the two mechanisms (Basho et al., 2007). For instance, Basho et al. (2007) used fMRI to show that imagined speech

evokes greater activation in multiple brain regions. In addition, Stark et al. (2017) conducted lesion symptom mapping (LSM) in subjects with aphasia. They discovered that those with poorer voiced speech were able to retain relatively good imagined speech. These findings indicate a dissociation of the cognitive processes responsible for the generation of voiced and imagined speech.

2.2 How BCI works

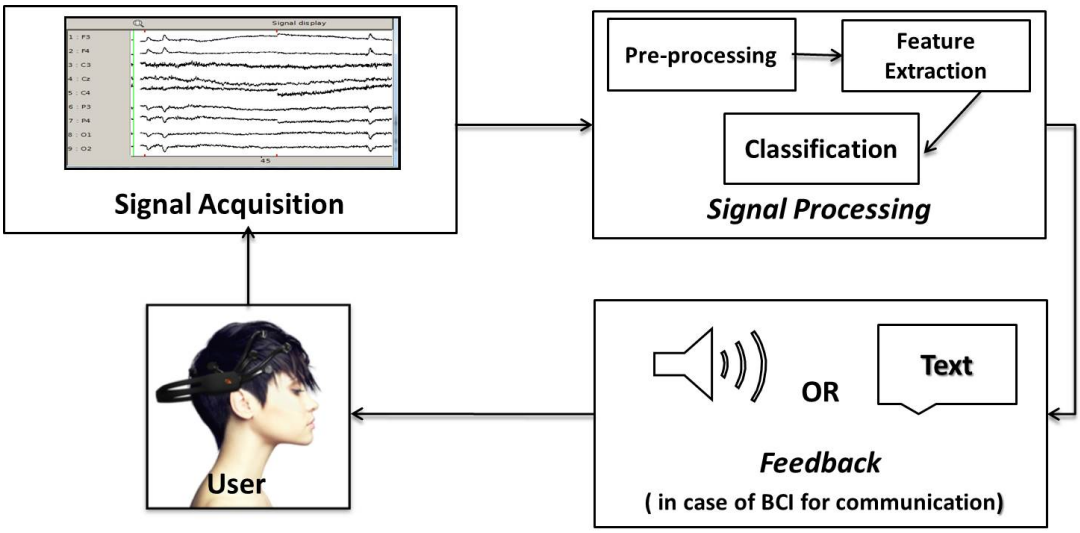

Fig. 2.3 Online brain computer interface cycle

Figure 2.3 shows the general framework for an online BCI system for communication. The collection of brain signals conveying informative neural features is the BCI input, while BCI outputs are letter or icon selections on a computer screen, a wheelchair control, or a neuro-prosthesis (Curran and Stokes, 2003; Vanacker et al., 2007). Every BCI system employs a unique algorithm to change its inputs into command signals so that the output device may be used in the intended manner. Van Gerven et al. (2009) describe five key areas of a BCI system, as follows:

1. **Signal acquisition**

   The user's brain signals are captured through a number of sensors during the

engagement in a specific mental activity. The BCI inputs are produced by the amplification and digitisation of the recorded brain signals.

2. **Pre-processing**

   Pre-processing algorithms are applied to maximise the signal-to-noise ratio by removing noise and artefacts in the brain signals that are gathered. Three types of pre-processing algorithms are commonly used: *artefact detection*, *spectral filtering*, and *spatial filtering*. In artefact detection, signals that are contaminated with dominant muscle activities are removed, such as signals related to eye or muscle movements. With spectral filtering, frequencies that are primarily involved in mental activities that are performed are captured, and any irrelevant frequencies are removed. Thus, spectral filters such as low-pass and band-pass algorithms are utilised to enhance the robustness of the signal by removing the noise from signals, such as so-called slow drifts and line noise. With spatial filtering, brain signals are combined from several electrodes in order to acquire information from several areas of the brain. An example of a spatial filtering tool is independent component analysis (ICA) (Hyvärinen and Oja, 2000); other examples of spatial filtering include channel referencing approaches such as common average referencing or the Laplacian filter (McFarland et al., 1997).

3. **Feature extraction**

   Feature extraction reduces the dimensions of data by extracting more informative features from the pre-processed signals. The extracted features may be temporal (time) features, spectral (frequency) features, or spectro-temporal (time-frequency) features.

   - **Time domain (TD) analysis** which stems from the desire to understand the signal in its original state, represents the first and most direct way to examine EEG signals. For example, TD analysis can be used to investigate alterations in EEG signals over time, including amplitude. The method also

includes traditional statistical measures, as in (Kumar et al., 2018), and cross-correlation, as in (Diwaker et al., 2016) work.

- **Frequency domain (FD) analysis** is employed to determine the frequencies that can be detected within the signals. At the same time, FD analysis is a viable way to illuminate the correlation between EEG frequency and amplitude, which then allows insights into the energy distributions within EEG signals. Fourier transform is commonly used to break down the time signals into sinusoids of several frequencies (Freeman and Quiroga, 2012).
- **Time-Frequency (TFD) analysis** of signals occurs when signals TD and FD are examined at the same time. Signals are represented in the present domain by a range of approaches including wavelet transform as discrete wavelet transform (González-Castañeda et al., 2017) and matched filters (DZmura et al., 2009).

4. **Classification**

   Classification algorithms are used with BCI to predict the user's intention based on the extracted features. Several classification algorithms are utilised for BCIs. According to Lotte et al. (2007), linear classifiers are the most widely used methods among BCIs, in particular linear discriminant analysis (LDA) algorithms. For modern classification algorithms, in Lotte et al. (2018) recent review, the authors provided several recommendations for which classification algorithms should be used for BCI designs. For small training datasets, for example, the authors recommend shrinkage linear discriminant analysis, transfer learning, Riemannian minimum distance to the mean classifiers, or random forest. Domain adaption is a useful method when the subjects are doing similar tasks, while deep neural networks are ineffective for BCI studies because of the limitation in data size. These modern classification algorithms still require further investigation, and several open questions remain to be answered (Lotte et al., 2018).

5. **Feedback**

   Feedback is another important step: the result of the prediction needs to be presented to the user. The different types of feedbacks may be summarised as visualisation (including light, colours, images, video, and motion animation; sonification or audification (including sound and music); haptic and tactile stimulus (including thermal, vibration, electrical, and micro-texture); and olfactory stimulation (including scent and pheromones).

2.3 BCI applications

After the signal has been classified to a particular class, the system can be used to link a particular command to the mental state in question, and this command is then sent to a given application. Two key categories of BCI applications are used. The first is the medical domain application category, as discussed in (Kubler et al., 2006; Wolpaw et al., 2002). Applications include disease prevention (such as with motion sickness), the finding and diagnosing of brain or sleep disorders, and the rehabilitation and restoration of brain function after events such as brain stroke. The second category of applications is in the non-medical domain (Blankertz et al., 2010). While BCIs are primarily intended for disabled people, they can also provide assistance to healthy people as well (Allison et al., 2007); examples include video games that use BCI.

Other applications are as follows:

- **Communication**

  Communication is a major goal for people with serious disabilities such as locked-in syndrome, where a person is entirely paralysed and cannot speak. BCI experts have attempted a number of different approaches for assistive communication. In this thesis unspoken speech is examined as a potential way of communication for disabled people. The first and best known approach for using BCI for communication is to control spellers (i.e. spelling programmes), either by using motor imagination–related rhythms or by using event-related potentials. Further

examples of the usage of BCI for communication are presented later in this chapter.

- **Mobility**

  An obvious focus of paralysed patients is the control of electrical wheelchairs (Rebsamen et al., 2010; Trambaiolli and Falk, 2018). The wheelchair can be controlled through reliable classifiable signals, which allows the paralysed patient to have a greater level of mobility than before.

- **Environmental control**

  A crucial application for EEG-based BCIs that allows quality-of-life (including independence), to be maintained for paralysed people. A number of environmental control systems of this type have been created in recent years, such as those discussed in (Ou et al., 2012; Zhang et al., 2017).

- **Virtual reality (VR) applications**

  VR systems are a new category of hybrid BCI applications. BCI uses a bio-signal as an input, and the VR component shows the output. At the moment, researchers focus on new input and output properties for BCI-VR with relation to various effects, purposes, and functions. An example of an effect would be to enhance feedback or to show results (Lotte et al., 2012), while purposes could include superior rehabilitation (Achanccaray et al., 2017) and psychological research (Fan et al., 2018). Lastly, functions could include new BCI-input modes that increase the level in which a person feels immersed in a virtual environment or bring about greater reliability in hybrid systems (Millán et al., 2010). VR has also been used throughout BCI training systems due to the technology's overall level of safety as well as the motivational elements involved. An example of this aspect may be seen in (Kryger et al., 2017) work, where a flight simulator was used to test BCI in the avionics context.

- **Neurorehabilitation**

  A benefit of BCI systems is their ability to present new therapy options for

patients with poor neuromuscular function caused by trauma or disease. Neurorehabilitation through BCI systems may be used for functional recovery in these patients and can be effective in boosting quality of life (Mak and Wolpaw, 2009). Previous studies suggest that a patient can be taught to increase brain activity in order to control motor functions in his or her body. This approach builds on the notion that with more normal brain activity comes increased central nervous system (CNS) function, thus building greater motor control. For example, early studies showed that stroke patients were able to regain control of certain brain activity patterns through this approach (Daly et al., 2006). Daly et al. (2006) studied EEG activity in stroke patients both before and after EEG-based neurorehabilitation. They found that, following the motor relearning intervention, there were changes in EEG features alongside a clearly positive change in motor function.

- **Gaming**

  Games that use BCI have a different level of interaction compared to those that use traditional keyboard and mouse controls. BCI control has a much larger potential for immersion, particularly for certain gaming methods. A number of studies have shown that BCI-controlled games, along side multimodal interactions, have complementary interaction benefits (Ferreira et al., 2014).

2.4 BCI technologies for signal acquisition

The technology used to measure brain activities can be divided into two main categories: *invasive* and *non-invasive*. *Invasive* BCI is implanted directly either into the brain or on the surface of the brain by neurosurgery. An example of *invasive* BCI is electrocorticography (ECoG). By contrast, *non-invasive* BCI scans brain activity from a user's scalp. This section describes each of the BCI technologies (*invasive and non-invasive*) and how they have been used as communication/speech tools in the literature. These applications can be classified into two types: BCI for speech recognition and

BCI for controlling spellers. BCI studies pertaining to speech recognition are explained in Chapter 3.

2.4.1 Invasive BCI

Electrocorticography (ECoG)

ECoG signals are recorded directly from the brain surface by using invasive BCI technology. The use of ECoG began in the 1950s, when it was used to localise epileptic seizures accurately before surgeries (Lotte et al., 2015). From the communication viewpoint, ECoG was used in the literature for unspoken and silent speech recognition (as presented in section 3.1).

2.4.2 Non-invasive

Functional magnetic resonance imaging (fMRI)

fMRI was discovered by Belliveau et al. (1991). It depends on measuring changes in local blood oxygenation levels during neuronal activation (Min et al., 2010). The benefits of fMRI include excellent spatial resolution and non-invasiveness. However, temporal resolution is limited (to $\sim$ 1 second), and the equipment is expensive. fMRI has been used in communication and speech studies (as explained in section 3.2).

Magnetoencephalography (MEG)

MEG measures the cortical magnetic fields created through electrical currents. It is a non-invasive methodology with excellent spatial and temporal resolutions. However, the equipment is very costly and requires the use of highly impractical isolation or shielding rooms because of the weak magnetic fields involved in the process. Moreover, MEG can be contaminated with artifacts specially movement of face muscles (Rampp and Stefan, 2007). MEG studies in speech are described in section 3.4.

Functional Near Infrared Spectroscopy (fNIRS)

fNIRS is a newer form of brain imaging that was first presented by Jobsis (1977). Although the majority of biological tissues (including skin, bones and muscles) do not absorb infrared light (620-700 nm), oxygenated haemoglobin does. This enables near-infrared light to be scattered through the scalp, skull and brain tissue and reach the outer cortex layers to record hemodynamic responses related to neuron behaviour. fNIRS has several advantages: it is not as affected by motion artefacts as other technologies, simple to set up in non-laboratory settings, more cost effective in comparison to fMRI and has better spatial resolution than EEG.

The neuroscience community has studied speech using fNIRS intensively. Research indicates that fNIRS may be useful in identifying the predominant setting for language function (Gallagher et al., 2007; Watanabe et al., 1998). More studies are presented in section 3.3.

2.5 Electroencephalography (EEG)

Electroencephalography is the recording of electrical activity of the human brain, as first introduced by Hans Berger in 1929 (Jasper, 1958a). Using EEG, the electrical signals created by action potentials of neurons in the brain are recorded through the scalp with the use of small metal electrodes. The advantages of EEG include: relatively low-cost, high temporal resolution, and non-invasiveness. The disadvantages include: low spatial resolution and non-stationarity in EEG signals (Min et al., 2010). EEG is employed in a number of BCI systems, and it was used in the research presented in this thesis.

The EEG records the brain's electrical activity from the scalp. EEG signals involve six main brain rhythms based on different signal frequency ranges: *delta* (1–4 Hz), *theta* (4–7 Hz), *alpha* (8–12 Hz), *mu* (8–13 Hz), *beta* (12–30 Hz), and *gamma*, at 25–100 Hz (Amiri et al., 2013). Each of these rhythms is enhanced under different conditions. For example, delta is more active among infants and in deep stages of sleep, while theta

Table 2.1 Comparison between non-invasive techniques for signal acquisition summarised from (Min et al., 2010), Res:Resolution.

Type	Signal source	Temporal Res	Spatial Res	Portable?	Cost (USD)
fMRI	Changes in Hemodynamic.	Low	High	No	> $1 million
MEG	Magnetic fields related to neuronal activity	High	Medium	No	$2–3 million
EEG	Electrical potentials related to cortical activity	High	Low	Yes	$200–$50 000

is more active during light sleep (normally in children under the age of 13). Alpha, which may be seen in the posterior regions of the head on both sides among normal adults, appears when a person is relaxing and disappears during periods of alertness (for example, when thinking or calculating). Beta, which can be seen on both sides of the brain in symmetrical distribution, is most clear in the front when a person is awake. The gamma rhythm can be detected in the somatosensory cortex when a person is highly engaged, while mu rhythm can be detected in the sensorimotor cortex when a person is performing a motor action.

High spatial and temporal resolution are both necessary to achieve the most effective brain measurement; other important factors are low cost, portability, and the ability to be used easily and non-invasively. Unfortunately, a BCI technology that combines all these features is not currently available. The properties of the various signal-acquisition methods are described in Table 2.1. Of those listed, the EEG technique is the most widely used for BCI. Its key advantages are high temporal resolution, relatively low price and ease of use. In addition, EEG is not an expensive approach when compared to MEG, or fMRI, all of which require costly equipment and trained professionals to be used correctly.

2.5.1 EEG acquisition device

EEG is used to record the electrical potentials created by the brain near its surface, so electrodes are placed on the scalp, or on the cortex itself. EEG recording systems have

four key parts: electrodes with conductive media, amplifiers with filters, an analogue to digital (A/D) converter, and a recording device. The electrodes interpret the signal taken from the head's surface, where amplifiers then move the microvolt signals into a range that can be reliably digitised. The converter then shifts the signals into digital form from analogue, before finally a personal computer or similar device is used to save and display the gathered data (Teplan et al., 2002).

In 1958, the International Federation of Electroencephalography and Clinical Neurophysiology (IFCN) established a standard for electrode placement known as the 10-20 electrode placement system (Jasper, 1958a). This approach created a universal standard for physical placement and naming for electrodes on the scalp. In this system, the head is divided into proportional distances from important skull landmarks such as the nasion, preauricular points, and inion in order to offer enough coverage across all of the brain's areas. Label 10-20 describes a proportional distance (in percent) from the ears and nose to where the electrode positions are established. Electrode positions are labelled based on the brain areas involved: F (frontal), C (central), T (temporal), P (posterior), and O (occipital). In addition to the letters, the left side of the head is assigned odd numbers, and the right side is assigned even numbers (Figure 2.4). The left and right sides are based on the perspective of the subject.

2.5.2 Neurophysiological signals in EEG used for BCI-based communication

The EEG activities commonly used in BCI may be categorised into three groups, depending on the component of interest: evoked related potentials (ERPs), slow cortical potentials (SCPs), and event-related de-synchronisation (ERD/ERS). Amiri et al. (2013) compared EEG activities based on four factors: accuracy, information rate, training time, and number of required EEG channels. They considered the ERPs brain activities to be the best because they are accurate, have a high information rate, and require less training time and using fewer EEG channels.

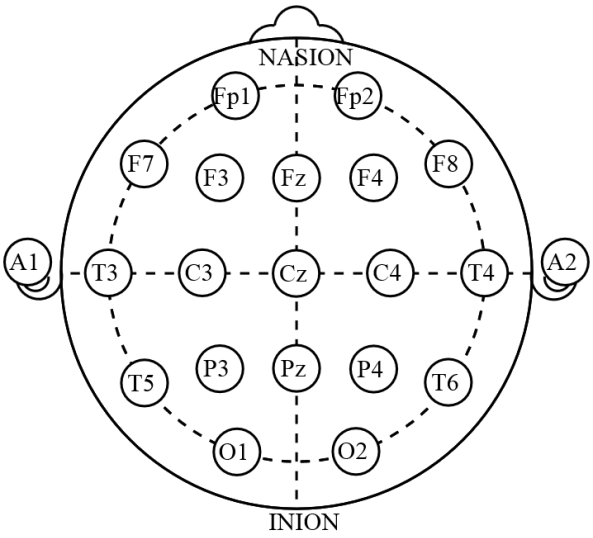

Fig. 2.4 Electrode locations of International 10-20 system for EEG (electroencephalography) recording (Wiki Commons-released to the public domain)

- **Evoked signals** are produced because of an external stimulus.
- **Spontaneous signals** are voluntarily produced by the user after an internal cognitive process, with no need to any external stimulations.

Evoked signals/ Evoked related potentials (ERPs)

ERP presents changes in the activity of neuronal populations, which can be detected at specific time delays after the appearance of a stimulus (Pfurtscheller and Da Silva, 1999). To enhance the signal-to-noise ratio, averaging techniques are used to detect these signals. Such activities have been used in the literature for two purposes: for communication and for understanding brain activities related to communication. Table 2.2 summarises the ERP features that have been investigated in the literature to date. As mentioned above, some of these activities are used for the direct control of communication systems such as P300, ErrPs, SSVEP, and N200.

- **P300**

  P300 was first described by Sutton et al. (1965). The amplitude and latency of P300 depends on a number of factors, including the inter-trial interval, the

probability of target stimulus, and the user's attention level. P300 can be detected from central electrodes such as Cz on a 10-20 international system. As a temporal pattern, the amplitude of P300 mostly falls in the range of 2–5 V, with a duration of 150–200 ms (Amiri et al., 2013). The P300-based speller is the most widely used application of P300 in BCI systems. Four paradigms are used for the design of these spellers: the row/column paradigm, the checkerboard paradigm, the single-character paradigm, and the region-based paradigm (Fazel-Rezai et al., 2012).

- **Error-Related Potentials (ErrPs)**

  To detect ErrPs, EEG signals are recorded from the frontocentral region, from Fz to Cz in 10-20 systems (Dehaene et al., 1994). In BCI systems, ErrPs is often used as an automatic error-detection mechanism. For example, (Dal Seno et al., 2010) used ErrPs as an automatic detection tool for a P300 speller. If ErrPs are elicited after the presentation of a letter chosen by the P300 speller, then the chosen letter will be cancelled.

- **Steady State Evoked Potential(SSVEP)**

  SSVEP usually includes the same fundamental frequency as the stimulus and a few harmonics of the fundamental frequency. For example, when a visual stimulus at a frequency ranging from 3.57 Hz is displayed, the brain generates electrical activity at in similar frequency, as well as at that frequency's harmonics (Amiri et al., 2013). Light-emitting diodes (LEDs) are usually used in BCI systems to generate SSEVP.

  SSEVP has been used to implement BCI spellers. For example, Chen et al. (2015) proposed an alphanumeric keyboard in which each key flicker targets a specific frequency and phase. The system can identify a target letter by detecting the elicited SSVEP frequency and phase in the user's EEG signal.

- **N200 using motion-onset visual response**

  N200 using motion-onset visual response is another example of ERPs. Hong et al.

(2009) used motion stimulus to generate the N200 signal in order to develop an N200 speller. They embedded motion stimuli into 6 x 6 matrixes that included alphabet buttons. In the study, motion was represented by vertical bars that appeared and moved leftwards for 140 ms at 200 ms intervals. The colour of the vertical bar, which could be red, green, blue, purple, yellow, or brown, was designated randomly in a protocol such that the colours of the six bars in the same row or the same column would be different from each other. The researchers compared the N200 speller with the P300 speller; their results showed that N200 could deliver performance comparable to P300 in terms of accuracy and needs less number of training trials.

- **Hybrid BCI paradigms**

  Any BCI system has certain disadvantages that prevent some users from adopting the system (Amiri et al., 2013; Wang et al., 2015). Recently, several researchers have tried to combine different BCI technologies (known as hybrid BCI) to develop a system suitable for a large number of users, or to include several tasks so that users could select a suitable BCI activity for each task (Amiri et al., 2013). This combination can be performed when each system has its own input (simultaneously) or when the output of one of the systems serves as the input of another system (sequentially).

  One example of a hybrid BCI system is the SSVEP/P300 BCI system developed by Wang et al. (2015). In this system, the researchers used changes in shape to elicit P300 and flickering to elicit SSVEP. They found that the general users' performance was better than that of the users who employed single-type BCI. The results also indicated that the use of shape change is preferable to the use of flashes for eliciting P300 with hybrid systems, because flashes cause a checkerboard phenomenon when presented along with flickering in the same system.

Table 2.2 Example of ERP patterns that are examined in the literature

ERP Features	**Description**
N100 or N1	A negative deflection observed when a stimulus is presented unexpectedly. The peak occurs between 90-200 ms after the stimulus.
P200 or P2	A positive deflection. The peak occurs approximately 100 ms to 250 ms after the stimulus. It is believed that the N1/P2 component of ERP may be characteristic of an individual's thrill-seeking behaviour.
N200 or N2	A negative deflection. The peak occurs approximately 200 ms after onset of the stimulus.
P300	P300 is a large and positive peak amplitude that,can be detected around 300 ms after the onset of a rare but relevant,stimulus.
N400	A negative wave, related inversely to the expectancy that a given word will form the end of a sentence. N400 is seen 300 ms to 600 ms after a stimulus. It was first reported with regard to semantic incongruity.
P600	An effect relevant to language processing. Sentences with syntactic errors, with a poor syntactic structure, or with a particularly complicated syntactic structure are associated with it.
SSVEP	SSVEP is evoked in the visual cortex as a response,when there is a repetitive stimulus with a constant frequency on the central,retina.
ErrP /ERN	Error-Related Potentials/ Error-Related Negativity is the average amplitude of the waveform at 50 to 100 ms after the onset of an aware error.

Spontaneous signals

- **Slow Cortical Potentials (SCPs)**

  SCPs is defined as slow changes in the voltage generated in the cortex, occurs over 0.51 seconds. SCPs is one of the lowest-frequency features recorded using EEG technologies. Negative SCPs are typically related to movement and any functions containing cortical activation, while positive SCPs are usually linked to a decrease in cortical activation (Wolpaw et al., 2000). Subjects can be trained to produce positive or negative SCPs, depending on the required task. In the literature, SCPs have been used to examine the potential to have communication systems for paralysed patients Birbaumer et al. (1999, 2003). SCPs are no longer used in BCI research, however, because of their lack of speed and their training time limitations.

- **Event-Related Desynchronization/Synchronization (ERD/ERS)**
  ERD/ERS relays on the recording of rhythmic activities over the sensorimotor cortex. In ERD/ERS, instead of having an external stimulus to generate a command as in ERPs, a user can voluntarily generate a command by controlling his or her brain activity, by imagining motor movements, or by any other activity (Rao and Scherer, 2010). ERPs have an advantage over ERDs, in that they are stable over time, while ERDs vary over time because the user can change the signals after receiving feedback. ERDs can be used for communication purposes, for example controlling BCI spellers based on hand and foot movement. The following is an overview of previous studies conducted on the use of BCI for speller control.

  - **Motor imagination to control BCI spellers**
    Guenther and Brumberg (2011) conducted a one-hour pilot study session with a single subject. The participant was asked to repeat three vowel sounds, with 20 repetitions of each sound: /AA/, /IY/, and /UW/. To represent each vowel, a different motor imagination action was linked with each sound: left-hand movement for /UW/, right-hand movement for /AA/, and foot pressing for /IY/. Limb imagery was used to ensure reliability and to obtain EEG responses. The recognised vowels were presented both visually and acoustically.

    In (Perdikis et al., 2014), the authors' main aim was to evaluate a BCI speller based on a motor-imaginary system called 'Braintree'. Left- and right-hand/ foot movement imagination was used for the selection of letters, and EMG signals represented the 'undo' command as an error-handling mechanism. To solve the problem of the limited options for brain actions that are used compared to the 26 letters in the English alphabet, the researchers used a language model as a data-compression technique to reduce the options available to the users. The letters were represented each time as a binary tree, so the model helped to redistribute the letters each time a letter was

selected. The model is called "prediction by partial matching". Braintree was tested on 16 end-users (6 disabled users and 10 able-bodied users). The evaluation had two phases: training and spelling. In the training phase, two motor-imaginary tasks were chosen, and the classifier was trained based on the selected data. In the spelling phase, the users were asked to spell four specific words, and there was no time constraint. All users were able to complete the spelling tasks; their performance was 1.7 characters per minute (cpm) on average, with 3.6 cpm being the maximum typing speed.

Dalbis et al. (2012), who developed a predictive speller controlled by motor-imaginary BCI actions, improved the speller's performance using a predictor. Three motor-imaginary actions were selected: (1) movement of the right and left hands, (2) movement of both hands for the alphabet, and (3) feet movements to undo actions. The alphabet was divided into three groups in the system interface; each time a user chose a letter, the groups were changed based on suggestions from the language model. Three subjects tested the system, achieving spelling rates of 3 cpm, 2.7 cpm, and 2 cpm.

2.6 Summary

BCI technologies are used as communication-assistive techniques in two different ways: to control spellers and for speech imagination. Figure 2.5 lists various BCI activities and their utilisation for communication purposes. For BCI-speller applications, non-invasive EEG technologies are used to elicit ERPs activities to control the spellers. In addition, some applications use motor-imaginary activities to control spellers. For example, hands and feet movements are often used for letter selection and for the undo function. For speech-imagination studies, both technologies (invasive and non-invasive) have been used to measure brain activities related to speech. For invasive BCI, several researchers have used ECoG in order to gain greater insights into the brain areas related to speech. ECoG can be used to retrieve accurate information in terms of time and spatial resolution, which is promising for the direct translation of brain activities into text or speech without having to average brain signals. Earlier studies used non-invasive BCI (such as EEG) to recognise a limited number of words, syllables, or letters. Some of these studies included other types of signals such as the aforementioned EMG and EOG to help in the recognition process.

This chapter has reviewed various approaches for the development of BCI applications. The basic building units of a BCI are the brain-signal measuring unit, the pre-processing and feature-extraction units, various classification algorithms, and experimental protocols. For non-invasive techniques for brain-signal gathering, EEG is the most widely used approach of those examined. The chapter has also presented a number of neuro-physiological signals that are commonly employed to drive EEG-based BCIs, with a focus on the patterns that appear alongside the techniques for communication purposes.

The focus of the research described in this thesis is to use EEG technology as an input technology using unspoken speech. The literature related to this area is summarised in the next chapter.

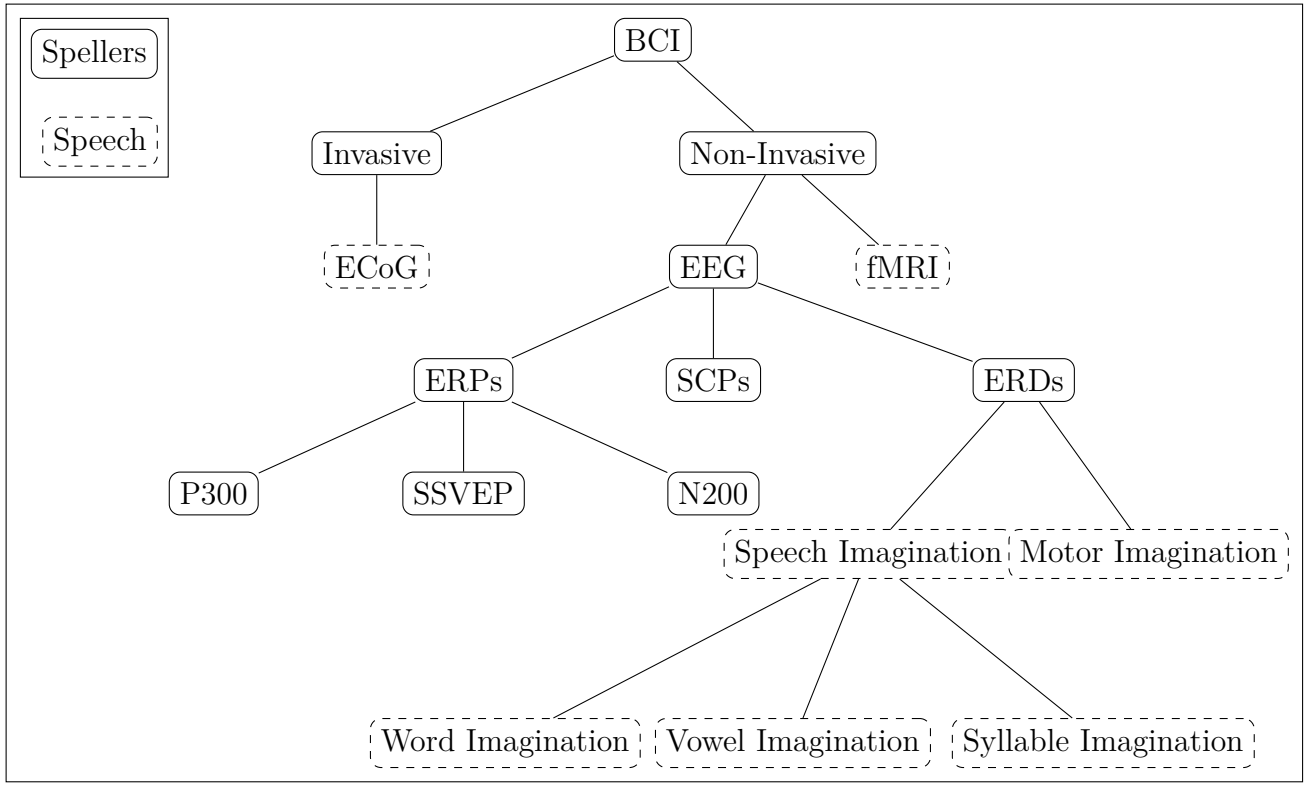

Fig. 2.5 Summary of BCI technologies and activities that have been examined for communication applications

Chapter 3

Brain Computer Interface for Unspoken Speech Recognition

Chapter 2 provided an overview of brain-computer interface (BCI) technologies as well as describing examples of BCI applications for communication purposes. The first category of BCI studies for communication is for controlling spellers (i.e. computer spelling devices). The second category is unspoken speech recognition (see section 2.6). Unspoken, imagined, or cvoiced speech can be defined as what occurs when subjects are asked to imagine the pronunciation of words as if they were pronouncing them aloud but without any articulatory movements.

The research reported in this thesis focuses on unspoken speech recognition from EEG signals. Researchers first began to express interest in understanding speech from EEG signals in 1997. (Suppes et al., 1997) was the first study of word recognition using EEG and MEG provided promising results about the speech-related information included in brain signals. Between 2006 and 2015, only a few studies examined the possibility of recognising unspoken speech using both invasive and non-invasive approaches. Some of these research studies were not conducted in English, however, while other works were restricted to a small number of subjects or a limited number of recognised parts of speech. Since 2017, the field has shown renewed interest in exploring different aspects of unspoken speech recognition in terms of the part of speech (vowel,

syllable, or word) that is examined, feature extraction techniques, and classification algorithms, in addition to various experimental factors that are often used to improve recognition rates.

Unspoken speech is very close to the natural way of communicating. The growing body of literature on the subject reflects the importance of examining this type of brain activity and of showing the potential of further improvements in recognition rates. This chapter reviews previous studies that have used BCI technologies (both invasive and non-invasive) for speech recognition. The chapter explains the methodologies the various authors followed to conduct their studies and reports on the results. The studies are categorised based on the sensors that were used to measure brain activities as well as by the different types of imagined speech that were performed. The end of the chapter includes a summary of the state of the art in this field as well as discussing study limitations. Part of this chapter stems from (AlSaleh et al., 2016) paper, which reviews studies on the subject between 2006 and 2016.

3.1 Invasive Electrocorticography (ECoG)

Section 2.4.1 of this thesis described invasive ECoG-BCI technology. In the literature there are several studies related to the understanding of spoken/unspoken speech from ECoG signals. Several studies have examined ECoG-derived brain patterns to decode audibly pronounced speech (Blakely et al., 2008; Kellis et al., 2010; Mugler et al., 2014; Zhang et al., 2012), understand the speech-production process in voiced and unspoken speech (Leuthardt et al., 2012; Lotte et al., 2015), speech perception, and feedback processing and understanding (Crone et al., 2001; Pasley et al., 2012).

For unspoken speech, in Guenther and Brumberg (2011) early study, two ECoG electrodes were implanted in one participant's motor cortex area, which the authors selected based on a previous study the same researchers had conducted in 2004. This area is assumed to be connected to articulatory movement. Following the presentation of a stimulus, the participant attempted to speak, and a synthesiser was used to

generate formants, vowels, and transitions. In this way, over the course of 25 sessions, the subject produced patterns and achieved a 70.00% success rate after 15 to 20 trials per session. Guenther and Brumberg (2011) results thus implied the clear possibility that direct BCI could be used for the direct synthesis of formants.

Pei et al. (2012) examined whether or not it was possible to determine the vowels and consonants of spoken and imagined words following visual and audial stimuli using ECoG signals. To answer this question, the authors examined four experimental conditions (visual stimuli/actual spoken, audial/actual, visual/imagined, and audial/imagined), with four possible vowels sounds (/ë/,/æ/, /i:/, or /u:/) and consonant pairs (/b_t/, /c_n/, /h_d/, /l_d/, /m_n/, /p_p/, /r_d/, /s_t/, or /t_n/) among 36 words. The findings showed that the brain areas activated during actual speech include the motor cortex, Broca's area, and the posterior superior temporal gyrus. In imagined speech, in contrast, two small foci in the temporal and frontal regions were found to have been activated. The results were promising, with classification accuracy rates of 55.00% in some cases among the four above-mentioned vowels.

Martin et al. (2016) recently conducted a study in which the authors used ECoG in the binary classification of words recorded in three different modes. First, the participants listened to a word, voicedly pronounced the word, or imagined the word. The words from each mode were classified independently. Six words were used: *"spoon"*, *"cowboy"*, *"battlefield"*, *"swimming"*, *"python"*, and *"telephone"*. These words were selected to have high variability in terms of semantic and acoustic features and numbers of syllables while still varying word length. The classification algorithm used in the study was improved using a nonlinear alignment algorithm to overcome the temporal variations the authors found between two trials (for the same word), which may have been caused by delays in the starting of a task or by differences in pronouncing/imagining words. The authors' proposed solution was to classify the high gamma features using a SVM with a dynamic time warping (DTW) kernel to align the features in a non-linear manner. The researchers found that their results were as expected. The classification rates among the listening and voiced speech tasks were high (listening: mean = 89.40%,

voiced speech: mean = 86.20%), while the classification rates of unvoiced speech averaged 57.70%.

Unlike many EEG-based brain-to-text systems that require the averaging of brain signals from multiple trials in order to have accurate silent-speech translation, ECoG-based systems work by using single-trial classification. The high signal-to-noise ratio of ECoG signals helps in better understanding the brain's speech-production mechanism. It should be noted that most of the participants in previous ECoG studies were seizure patients who used the ECoG electrodes primarily to localise their epileptic seizures. Typically, the use of ECoG for unspoken speech recognition is limited because of the invasiveness of the technology.

3.2 Functional Magnetic Resonance imaging (fMRI)

Belliveau et al. (1991) were the first to discover and report on fMRI. The approach depends on measuring changes in local blood oxygenation levels during neuronal activation (Min et al., 2010). Since that time, the technique has been widely applied in communication and speech studies. For example, Naci et al. (2013) sought to answer the question of whether fMRI could be used to decode binary (i.e. *"yes"*/*"no"*) answers by conducting an experiment using 15 participants with no neurological disorders. The participants were asked if they had any sisters or brothers; 90.00% of their answers were decoded correctly within three minutes of fMRI scanning, which demonstrated that the approach could provide an accurate and relatively fast tool for communication.

Yang et al. (2014) conducted a further study, both to classify between *"yes"* and *"no"* answers and to determine if the hemodynamic spatial patterns they examined included sufficient information to classify between words. The researchers first examined the whole brain during the voiced speech involved in answering binary questions. The brain areas that were activated included the right superior temporal gyrus, the left supra-marginal gyrus, and the left middle frontal gyrus. The authors then examined these areas in a second experiment to decode the unspoken answers (*"yes"*/*"no"*),

which yielded rates of 82.50%, 77.50%, and 79.50%, respectively, for the different brain areas. The researchers also noted that the decoding of the unspoken answers was similar, regardless of the subjects' intention to answer honestly or dishonestly.

In addition to word classification, Huth et al. (2016) used a word map overlaid on the brain to investigate semantics. To do this, seven native English-speaking subjects were asked to listen to 10,470 words of narrative stories while fMRI scans were conducted. The words in the stories were clustered into 12 groups using K-mean clustering, and each category was inspected and labelled by hand. Because the results of the study showed consistency in the organisation of words among users, the authors considered that this similarity was due to the users' backgrounds. Huth et al. (2016) thus suggested that future work in this area should necessarily include subjects from different backgrounds and languages.

In general, because of bulkiness, immobility, and a slow time response, the use of fMRI as a communication system may not be feasible in daily life, although its good spatial resolution does help to understand the brain activities associated with cvoiced and voiced speech production.

3.3 Functional Near-Infrared Spectroscopy (fNIRS)

As was explained in section 2.4.2, fNIRS has several advantages to be used in speech studies. In Gallagher et al. (2007) study, a fully locked-in patient was asked to respond to *"yes"* or *"no"* questions by conducting mental repetition of the answer for 25 s. The study had an offline average classification accuracy of ∼76%. Likewise, Hwang et al. (2016) classified between the voiced speech of *"yes"* versus *"no"*.

Other research has been conducted to discriminate between different speech modes. Herff et al. (2012) asked five subjects to produce audible, silently uttered, and imagined speech or to not speak at all. The subjects' fNIRS signals were recorded during every speaking mode. Using SVM and mutual information, between 69% to 88% classification accuracies were obtained. The three modes of speech showed an average classification

accuracy of 61%. The researchers concluded that the results are comparable to motor imaginary results using fNIRS.

To propose natural way of communication using BCI, Herff et al. (2013) examined fNIRS for automatic detection of subject's intention to speak. The researchers achieved this goal through the detection of asynchronous speech activity, which is a highly natural type of communication.

3.4 Magnetencephalography (MEG)

Several studies in the literature have used MEG to examine the speech process. For example, Houde et al. (2002) examined the interaction between speech production and perception. While Tervaniemi et al. (1999) measured the differences in the processing of musical and phonetic sounds in the human auditory cortex. Heinks-Maldonado et al. (2006) used MEG to prove the accuracy of a forward model of speech production, which entailed a comparison of the feedback during speech with internal predictions. Guimaraes et al. (2007) presented single-trial MEG classification during visual and audio word stimulation.Wang et al. (2017) recently presented a study of the voiced speech of five commonly used phrases.

3.5 Electroencephalogram (EEG)

Several studies have used EEG to explore the possibility of reading what people are thinking about. Experiments with speech imagination can be divided into three types, based on the imagined part: word imagination, syllable imagination, and vowel imagination. Brigham and Kumar (2010b), who used imagined speech as a user-authentication technique, argued that the use of speech imagery is more convenient and intuitive for users than motor imagery or any other type of mental activity, although they did note that additional research and practice on speech imagery are both required

to establish the most appropriate methods for generating discriminative EEG signals without any voiced actions.

Because of its portability and cheapness, EEG shows the greatest potential among other modalities for use as a communication system for daily life. In particular, advances in sensor technology are likely to lead to wireless, dry, and less expensive EEG sensors, although EEG's low signal-to-noise ratio and the inherent non-stationary quality of EEG signals make speech recognition a challenge. Advances in signal-processing algorithms are also required in order to yield more robust and accurate communication systems based on EEG. Building on this idea, an overview of previous research studies on imagined speech recognition is presented below; this section also includes further details on the signal processing approaches used in previous studies, ordered by type of word stimuli (words, syllables, and vowels) as well as by study date.

3.5.1 Word imagination

Wester (2006) was among the first to examine the possibility of imagined speech recognition using EEG. In addition to examining unspoken speech, the author also studied whispering, silent speech, and silent mumbling. The trials for each word were recorded in blocks, where the word was first presented, then the trials for the word were conducted. The features the author examined included windowed short time Fourier coefficients, delta coefficients, and delta mean coefficients. For classification, the author used the hidden markov model, where the number of gaussian distributions varied based on classification accuracy.

First, the author compared different modalities and found that the results were four to five times higher than chance, with the results from unspoken speech poorer than other modalities but still comparable across comparisons. Second, the author discussed, for one subject, the effect of understanding of the meaning of a word upon recognition by considering the reactions of non-native English speakers to words that they could not understand. For this approach, the accuracy for these words was higher than that for others, which was a result of the use of pronunciation rather than picture

imagination. Finally, the author considered which parts of the brain are the most accurate at recognition, with a focus on the Homunculus, Broca and Wernicke areas. The results from this investigation showed that accuracy was slightly better when classification was conducted using all the electrodes rather than just those focussed on these areas. The results showed that these brain areas, in addition to the homunculus, are the most important for recognition.

Porbadnigk et al. (2009) attempted to prove that speech imagination can be recognised effectively if the spoken words are in blocks (i.e. a sequence of words), where the words are separated using eye blinks. The study showed a relationship between word order and recognition rate. More specifically, recording unspoken speech in blocks allowed the users to concentrate more (as they reported) and to provide signals with less noise and consequently better recognition rates. The researchers did express concern that the recognition was affected by temporal artefacts due to the repetition of words. The researchers concluded that the existence of these artefacts did not mean that unspoken speech was not there but that they needed to change their experiment to yield better recognition data. The experiment involved five words ("alpha", *"bravo"*, *"charlie"*, *"delta"*, *"echo"*), 21 subjects, and 16 channels. The average classification accuracy between the five words was 45.95%.

González-Castañeda et al. (2017) main objective was to assess new algorithms for the classification of unspoken words, including sonification and textification. Sonification is a technique in which EEG signals are processed in the audio domain; the technique thus has the advantage of permitting the use of automatic speech recognition technology, which results in clearer patterns than were present in the original signal as well as the existence of more features from which automatic classifications can be conferred. Following sonification, features were computed using energy computed from the decomposition levels of each discrete wavelet transform (DWT). The ultimate goal of textification is not in fact the creation of conventional text sentences; rather, the technique is used to transform EEG signals into textual coding so that text mining techniques may be applied. The overall approach is informed by 'bump extraction',

which promotes the identification of higher-energy time and frequency zones. The experiment involved the imagination of five Spanish words: *"arriba (up)"*, *"abajo (down)"*, *"izquierda (left)"*, *"derecha (right)"*, and *"seleccionar (select)"*.

Similar to (Porbadnigk et al., 2009; Wester, 2006), the trials for each word were recorded in blocks, with 33 trials for each word. Emotiv EEG device was used for the recording, and all the 14 channels were considered in processing EEG signals. The average accuracy using the original EEG signals was 58.41%, while the rates for the sonified and textified EEG signals were 63.82% and 83.34%, respectively. Thus, the sonification and textification of EEG clearly led to significant improvements in these experiments.

Torres-García et al. (2016) main aim was to provide a channel selection method for imagined speech classification using the same dataset described in (González-Castañeda et al., 2017). The researchers proposed a method of channel selection. The proposed algorithm was successfully, providing 68.18% when using seven channels compared to using all channels, which provided 70% average accuracy.

Sereshkeh et al. (2017b), who claim to be the first to suggest the online EEG classification of unspoken speech, had two main objectives in their work. In the first test, the authors aimed to distinguish between a ten-second episode of imagined repetitions of the word *"no"* and a ten-second period of rest. In the second test, the goal was to differentiate between two ten-second periods of *"yes"* and *"no"*. The study sample consisted of 12 subjects; they each participated in four sessions, two of which were offline training sessions. Where two sessions were related to the same task, the first session was used for training, while the second was the online session. Overall, 140 trials were recorded for each training class, and EEG signals were recorded from 64 electrodes. Both statistics on DWT coefficients and autoregressive coefficients AR were used for feature extraction and SVM was used for classification, wherein the best ten features were chosen using a correlation-based filter. An average accuracy of 75.90% was obtained for participants in online sessions with *"no"* versus with rest. The average accuracy was 69.30% for the *"yes"* versus *"no"* sessions.

Sereshkeh et al. (2017b) also examined issues beyond classification. First, they considered the significance of brain regions in view of the data they obtained from the motor regions, which included signals linked to possibly motor movements or potential confusion arising from sub-vocalisations. The authors concluded that the data was not significant for either classification problems. Second, the authors found that average accuracy trended upwards in accordance with increased trial duration. Finally, for the "no" versus rest classification, the authors found that the cross-over point was reached at three seconds.

Rezazadeh Sereshkeh et al. (2017), who employed the same dataset as Sereshkeh et al. (2017b), used a multilayer perceptron (MLP) artificial neural network (ANN) with statistics on DWT as features to classify all three pairing combinations of *"yes"*, *"no"*, and rest as well as tertiary classification. All subjects surpassed the average accuracy of 75.70% in the case of unspoken word trials versus rest, and *"yes"* versus *"no"*. In the tertiary classification, all participants surpassed the chance level, with an average accuracy of 63.20%. In comparison the other classification techniques, the MLP network provided higher classification accuracy. Specifically, the authors examined SVM with polynomial kernel and linear, k-nearest-neighbours (KNN), LDA, and naive Bayes (NB).

Balaji et al. (2017) recently proposed a bilingual interpretation and decision-making approach. Five subjects who were proficient in Hindi (as their first language) and English (as their second language) were involved in the study. During the experiment, 10 obvious questions were asked randomly in both languages; the subjects had their eyes closed throughout the experiment. The questions were audibly recorded, and the subjects had 10 seconds to answer each question by thinking about the semantics of the answer in Hindi or English. The researchers found that the EEG data they acquired involved information about the language, decision-making, and the answers. After pre-processing the EEG data and applying dimensionality reduction, the authors further reduced the data dimensionality using principal component analysis (PCA). The data from all subjects was used to train the classification algorithms: SVM, random

forest (RF), AdaBoost, and ANNs. The results showed that the ANN classifier had the best classification accuracy: 85.20% and 92.18% for decision and language classification, respectively. The classification accuracy of bilingual speech was 75.38%.

In Nguyen et al. (2017) study, their main aims were to examine several issues related to unspoken word recognition and to propose a classification algorithm based on using Riemannian manifold features. The recording was performed in different sessions, where each session had a different classification aim. The total number of subjects was 15, although each subject participated in up to three sessions. The sessions were used to classify between vowels (*"/a/"*, *"/i/"*, and *"/u/"*), short words (*"in"*, *"out"*, and *"up"*), long words ("cooperate" and "independent"), and a short word (such as *"in"*) versus a long word (such as *"cooperate"*). Not all the subjects participated in the four experiments: in each experiment, five to six subjects were involved.

The subjects were trained to perform the imagination in specific time lengths by a beeping sound. In the case of vowels and short words, the time between beeps was 1 s, whereas the time between long words and between short words and long words was 1.4 s. The subjects then had to repeat the word three times without the beeping sound, as they had already been trained on the required length. Each one of the three imaginations was considered to be one trial. For feature extraction, the researchers proposed the use of the Riemannian manifold features framework and a relevance vector machine for classification. The researchers also involved the subjects in experiments that were relatively similar to previous works on vowel classification. The four experiments resulted in 49.00%, 50.10%, 66.20%, and 80% for vowels, short words, long words, and short words versus long words, respectively.

Qureshi et al. (2018) recently proposed an algorithm to classify EEG data related to the imagination of five English words: *"go"*, *"back"*, *"left"*, *"right"*, and *"stop"*. In EEG data recording, one second after the target audio stimulus began, a double-beep sound started, with a 500-ms gap between each beep. The purpose of the beeping sounds was to establish time cues, such that the subjects could employ a uniform time frame in which to imagine words. Subsequently, 500 ms after the second beep, the

participants were told to re-imagine the word they had heard two seconds earlier at the start of the process.

Covariance-based connectivity, and maximum cross-correlation were selected as the extraction methods. For classification, a sigmoid activation function-based extreme learning machine (ELM) was compared with linear and non-linear SVMs, RF, and KNNs. In terms of classification accuracy and computation time, ELM outperformed the other classifiers. The maximum average accuracy for the classification of the five words was 40.30%.

3.5.2 Syllable imagination

DZmura et al. (2009) attempted to distinguish linguistic content from brain waves by determining the brain signature of specific linguistic content; in this way, the signatures are indicated as the differences between alpha, beta, and theta brain patterns. During this experiment, the subjects were asked to imagine speech containing the two syllables /ba/ and /ka/ in three different rhythms (six conditions) without any effort or muscle movement. Two analytical techniques were applied. First, the authors computed matched filter classification using envelopes for the three frequency bands (alpha, beta, and theta) at each electrode to generate a series of classification accuracies for the three frequency bands for each subject. Second, they calculated the power spectral density per condition for every frequency band; the authors then showed the power distribution over the electrodes for each spectrographic representation. Using these approaches, the researchers concluded that speech imagination recognition could be attained for each subject separately but that averaging across subjects did not yield accurate information; their results showed that active frequency bands and electrodes were different between subjects and between conditions. The authors did not provide classification accuracy data in their study.

The main aim of Brigham and Kumar (2010a) work was to use imagined speech for subject identification and authentication, with six subjects. In addition to testing a proposed signal analysis method on EEG signals related to the subjects' imagined

speech, the researchers also examined a database of EEG signals related to visual evoked potentials (VEPs) for use in subject identification, with 120 subjects. During the imagined speech testing, the subjects were asked to imagine the speech of two syllables (/ba/ and /ka/) at different rhythms. The researchers argued that the use of syllables in imagination instead of full words would avoid semantic effects on brain signals. AR coefficients were used for feature extraction, while SVM was used for feature classification. The authors' signal processing method showed a high level of subject identification accuracy (99.76% for syllable imagination and 98.96% for VEPs). They did note, however, that this accuracy decreased when further sessions were recorded, which may have been due to participant fatigue.

3.5.3 Vowel imagination

An early study by DaSalla et al. (2009) aimed to distinguish between /a/ and /u/ and no imagination as controlling task. Common spatial patterns were used as classification features. Three healthy subjects were instructed to imagine the mouth movement during the imagination of each vowel. Non-linear support vector machine was used as classification algorithm. The classification accuracy were between 68% to 78%.

Chia et al. (2011) conducted a study to discriminate between phonemes with different vocal articulations (jaw, tongue, nasal, lips, and fricative). The authors collected the data from five subjects using an EEG device. The five classes were compared in pair-wise classification between the classes and against the time in which no imagination occurred. Spectrographic data was classified using two different classifiers: naive Bayes and LDA classification algorithms. The pair-wise classification results showed 80% or above classification accuracy. The researchers examined the data that had been recorded on different days; as expected, the data recorded on the same day provided better results.

Matsumoto and Hori (2014) focussed on vowels (/a/, /i/, /u/, /e/, and /o/) because they targeted the Japanese language, where the structure of the syllables consists of one vowel and one consonant. They examined the differences between two classification

algorithms – an RVM with a Gaussian kernel (RVM-G) – and compared the results with those generated by SVM with a Gaussian kernel (SVM-G), described in (Matsumoto and Hori, 2013). The purpose was to reduce calculation costs while using 19 channels, common spatial patterns (CSPs) filtering, and adaptive collection (AC). The findings showed no differences between RVM and SVM in terms of classification accuracy, which were both in the 77%–79% range. The calculation costs of RVM are also higher, and the technique requires more training data to provide strong results.

Sarmiento et al. (2014) focussed on distinguishing between the mental state imagination of open, mid, and closed vowels without the imagination of the articulator movements. Twenty-one electrodes were placed over subjects' Wernicke's and Broca's areas, as these areas are related to speech. In the pre-processing stage, the differences between articulation modes were calculated based on time domain analysis and by applying periodograms using two fixed factors: the stimulus applied to the subject and the position of the 21 electrodes. Power spectral analysis was applied to detect signals that were immersed in noise by considering only the signals between the ranges of 2 to 16 Hz. Finally, the classification was conducted with a non-linear SVM, which resulted in recognition rates between 84% and 94%.

The main purpose of Idrees and Farooq (2016) study was to examine the viability of using features extracted from the beta, delta, and theta rhythms of EEG to classify imagined two English vowel sounds: /a/ and /u/, along with a 'rest or no action' condition. The imagination process involved subjects imagining the mouth movement associated with each vowel. Three subjects were involved in the experiment. EEG electrodes were placed on the motor cortex, and a 0–500 ms time frame was given for each imagination task. Features in the 0–8 Hz and 16–32 Hz range were obtained using wavelet decomposition. The features involved total energy and the energy's waveform length of the approximate and detailed coefficients. The average accuracy of pairwise classification was 65–82.50%, while that for the combination of tasks was between 81.25% and 98.75%.

Jahangiri and Sepulveda (2017) study aimed to classify between five imagined words that were abbreviated to phonemes. The word *"back"* was abbreviated to *"BA"*, *"forward"* to *"FO"*, *"left"* to *"LE"*, and *"right"* to *"RY"*. The study involved ten subjects aged 22 to 70, eight of whom were neurologically healthy, one of whom had dyslexia, and one of whom had autism. Discrete Gabor transform was applied for feature extraction. The average pair-wise classification accuracies using LDA were between 72% and 88%. In addition of conducting pair-word classification, the researchers examined the contribution of the well-known frequency bands in the classification. The findings showed that 12% and 31% of the important features were connected to the alpha and beta bands, respectively, which is similar to the motor imagination findings mentioned in the literature; 57% of the features were linked to high gamma brain activities.

Jahangiri et al. (2018) main aim was to compare the classification accuracy between the same four imaginary speech phonemics used in (Jahangiri and Sepulveda, 2017) ("BA", "FO", "LE", and "RY") with the accuracy of classifying four classes of motor imaginary directions (left hand movement, right hand movement, left foot movement, and right foot movement), where the phonemic also represents directions. The two experiments were conducted using a similar design for both tasks in order to achieve similar comparisons with the four subjects. The stimuli in both experiments were presented as arrows in random order. For EEG data acquisition, the authors used Enobio with 20 electrodes. Discrete Gabor transform was applied for feature extraction, and LDA was used as the classification algorithm. In both experiments, only 10 trials were recorded from each task (eight training and two testing). The researchers stated that, while their aim was not to increase classification accuracy (since both phonemics and motor imagination displayed a high level of accuracy). The average obtained classification accuracy was 82.50% for imagined speech and 77.20% for motor imagery. The paper's main discussion was about the dominant frequencies in both activities. In speech imagination, high gamma activities were found to have increased, which was supported by Jahangiri and Sepulveda (2017) previous findings.

3.6 Discussion on current EEG-Based studies

This review chapter discussed prior research related to the use of BCI technologies and how these technologies have been used to understand speech production and recognition processes. The discussion in this section emphasises those studies that have utilised EEG signals (since it is the signal of interest in this research), broken down into three parts: (1) the type of stimuli used in the experiments and tasks design, (2) the feature extraction algorithms, and (3) the classification accuracy that each study achieved. Table 3.1 provides a summary of EEG studies for unspoken speech recognition.

3.6.1 Stimuli and tasks design

In previous studies' stimuli selection has been based primarily on language aspects. For example, Brigham and Kumar (2010a) and DZmura et al. (2009) used syllables (because syllables lack semantic meaning). While Sarmiento et al. (2014) used the imagination of mouth shape of different vowels; Jahangiri and Sepulveda (2017) used the phonemic representation of words. Where the researchers assumed that these acoustic differences will resulted in variations in the brains activities due to the differences in the face muscles involved in the imagination of these words. In (Nguyen et al., 2017), the differences in the complexity between the words were considered where short words (two letters) and complex long words were compared.

Balaji et al. (2017) added another two parameters in order to enhance the classification of unspoken words: the decision-making that occurs while thinking of answers to questions, and a bilingual experimental paradigm. It is also worth noting that, in terms of languages, not all previous studies have been conducted in English. For example, González-Castañeda et al. (2017) used Spanish stimuli, while Matsumoto and Hori (2014) used Japanese stimuli as the basis of their studies.

Studies related to the recognition of unspoken words using EEG typically divide the design of tasks into three categories, depending on the length and repetition of the speech task. The first category is block recording, in which the participant is informed

before each block about the word that he or she should imagine (González-Castañeda et al., 2017; Porbadnigk et al., 2009; Wester, 2006). The participant is then asked to repeat the same word for a specific number of trials. The trials are separated using either eye blinks, as in (Porbadnigk et al., 2009), or mouse clicks, as in (González-Castañeda et al., 2017). In addition to the type of separation techniques that are employed, the number of trials included in each block for every word varies across studies; for example, (Porbadnigk et al., 2009) used 45 trials, while (González-Castañeda et al., 2017) used 33 trials.

The second category involves randomly presenting a written or audio-recorded word, syllable, or vowel to the participant. After the stimulus disappears, the imagination should be performed once within a specific time frame, which varies between studies. For English vowel imagination, as in (Yoshimura et al., 2011), the time was two seconds, whereas for Japanese vowel imagination, as in (Matsumoto and Hori, 2014), it was one second. In (DZmura et al., 2009), the participants were instructed to imagine syllables within a different time period on the basis of the required rhythm. The presentation of the stimuli was repeated randomly.

The third category was presented for the online recognition of *"yes"* and *"no"* (Rezazadeh Sereshkeh et al., 2017; Sereshkeh et al., 2017b). The stimuli were a set of questions, and the participants had to answer the questions by imagining either *"yes"* or *"no"*. Each trial lasted 10 seconds, and the participants repeated the imagination for an unlimited number of times.

3.6.2 Feature extraction

The approaches that are typically followed for feature extraction can be divided into: time domain, frequency domain, time-frequency domain, and spatial features. For time-domain features, Min et al. (2016) provided classifiers with statistics of EEG time series as features. In Brigham and Kumar (2010b); Sereshkeh et al. (2017b) autoregressive (AR) coefficients was used as another feature in addition to time-frequency features. In a more recent study, Qureshi et al. (2018) used covariance-based features (e.g. speech

area channels versus the rest of the channels) and maximum linear cross-correlation coefficients as features for recognising speech imagination. Balaji et al. (2017) used reduced-dimensionality EEG signals to classify between *"yes"* and *"no"* to questions that were answered in English and Hindi.

Only a few early studies used frequency domain features, Wester (2006) used windowed short time fourier coefficients as a feature extraction method. Chia et al. (2011) used spectrographic representation of EEG data as a classification method. Sarmiento et al. (2014) used power-spectral density to extract the frequency of interest from EEG signals.

Most studies have used time-frequency features. For example, several studies have investigated wavelet-based features, which can be extracted using techniques such as matched filters (DZmura et al., 2009); discrete wavelet transform (DWT) (González-Castañeda et al., 2017; Idrees and Farooq, 2016; Sereshkeh et al., 2017b); wavelet packet decomposition (Abdallah et al., 2017); or discrete Gabor transform (Jahangiri et al., 2018). For spatial filters, DaSalla et al. (2009) and Matsumoto and Hori (2014) used common spatial patterns, which are widely used in motor imagery classification studies.

Far from the main categories of features, González-Castañeda et al. (2017) transformed EEG patterns related to unspoken speech into audio (i.e. sonification) and text (textification) before feature extraction.

3.6.3 Classification accuracies

A final important point to discuss is classification results. Large variations exist in the literature in terms of number and type of stimuli, the number of subjects, and the feature extraction methods used to distinguish EEG patterns related to unspoken speech. The EEG devices used for signal acquisition also vary, with some studies using 64-channel devices, while others use relatively cheap wireless EEG machinery. In general, studies based on vowel imagination (i.e. the imagined pronunciation of a vowel or mouth shape) or the imagination of the phonemic representation of a word

achieved better multi-class categorisation than other studies. As the pronunciation differences and phonological characteristics can be captured in imagined speech similar to audible speech (see section 2.1.2 for more discussions).

On of the hypothesis that described the importance of the articulation differences is the flexible abstraction hypothesis (Oppenheim and Dell, 2010). The researchers proposed that examples where imagined speech seemed to have phonological characteristics may have been the result of subjects employing a variety of imagined speech that utilised articulation to a large extent. According to this hypothesis, although imagined speech may lack articulatory representations, it could exhibit lower-level articulatory planning when speakers employ silent articulation. This articulation adds extra information to help in distinguishing the imagined stimuli.

Table 3.1 Summary of EEG-based studies in imagined speech recognition

Study	Stimuli	Task Design	Subjects	Brain Area/Number of Electrodes	Features	Classifiers	Average Performance
(Wester, 2006)	Different vocabularies groups, different modalities: i.e. whispering, silent speech, silent mumbling and unspoken speech	block recording	5 subjects	Homunculus, Broca, and Wernickes area	Windowed short fourier transform	HMM	Higher than chance level and different based on experimental parameters
(Porbadnigk et al., 2009)	Five English words	Block recording	5 subjects	Orofacial motor cortex	–	HMM	45.50%
(González-Castañeda et al., 2017)	Five Spanish words	block recording	27 subjects	14 channels	DWT on raw and sonified EEG, and textification	RF, and NB	Raw EEG: 58.41% sonification: 63.82% textification: 83.34%
(Sereshkeh et al., 2017b)	Two words: “yes”, “no”, and rest time	Random trials where each was 10 seconds of episode of imagination	12 subjects	64 electrodes	Statistics on DWT and autoregressive coefficients	SVM	“no” vs rest: 75.90%, “no” vs “yes”: 69.30%
(Sereshkeh et al., 2017b)	Two words: “yes”, “no”, and rest time	Random trials where each was 10 seconds of episode of imagination	12 subjects	64 electrodes	Statistics of DWT	ANN, LDA, SVM, and KNN	“no” versus rest: 74.93%, “yes” versus rest: 73.19% “no” vs “yes” vs rest: 53.18%
(Balaji et al., 2017)	“yes” and “no” in English and Hindi	Answering random questions	5 subjects	32 electrodes	Raw EEG	SVM, RF, AdaBoost, and ANNs	Decision: 85.20% Language: 92.18% Multiclass: 75.3%
(Nguyen et al., 2017)	three short words, two long words, and three vowels	Randomly presented words where each group of words recorded separately for each experiment aim	15 subjects	64 electrodes	Riemannian manifold based features	Relevance Vector Machine	Vowels:49.00% Short words: 50.10% Long words: 66.20% Short words vs long words: 80%
(Qureshi et al., 2018)	Five English words	Randomly presented words	8 subjects	64 electrodes	Covariance-based connectivity, and maximum cross-correlation	ELM, SVMs, RF, and KNNs	40.30%
(DZmura et al., 2009)	Two syllables /ba/ and /ka/	Random trials and imagination in three different rhythms	4 subjects	128 electrodes	Spectrographic representation of the data	Matched filter classification	not provided
(Brigham and Kumar, 2010a)	Two syllables /ba/ and /ka/	Random trials and imagination in three different rhythms for the aim of subject identification	6 subjects	128 electrodes	autoregressive coefficients	SVM	Subject identification:99.76%
(DaSalla et al., 2009)	/a/ and /u/ and no imagination	Random trials for mouth movement imagination	3 subjects	64 electrodes	Common spatial patterns	Non-linear SVM	between 68% to 78%
(Chia et al., 2011)	10 English-language phonemes (five classes each class of two phonemes)	Imagine mouth movement for each phoneme	5 subjects	52 electrodes	Spectographic data	NB and LDA	pair-wise classification: 80%
(Matsumoto and Hori, 2014)	/a/, /i/, /u/, /e/, and /o/ in Japanese	Random trials	5 subjects	19 electrodes	Common spatial patterns and adaptive collection	RVM and SVM	between 77%–79%
(Sarmiento et al., 2014)	Distinguish between open, mid, and closed vowels	Continuous imagination of each vowel separately	5 subjects	21 electrodes over Wernicke and Broca areas	Power spectral analysis	Non-linear SVM	Between 84% and 94%
(Idrees and Farooq, 2016)	Two English vowels: /a/ and /u/, and no imagination	Imagining mouth movement for randomly presented vowel	3 subjects	Motor cortex	Wavelet decomposition	LDA	Pairwise classification: 65–82.50% Ternary classification: 81.25%-98.75%
(Jahangiri and Sepulveda, 2017)	Phonemic representation of five English words	Randomly presented words	10 subjects	64 electrodes	Discrete Gabor transform	LDA	Pairwise classification:72%-88%
(Jahangiri et al., 2018)	Phonemic representation of five English words	Randomly presented words	4 subjects	20 electrodes	Discrete Gabor transform	LDA	82.50%

3.7 Summary

The review presented in this chapter consists of two parts. The chapter first presented studies that have demonstrated how different BCI technologies may be used to understand speech production and recognition. The second part of the chapter examined studies that have utilised EEG signals for unspoken speech recognition.

Based on this review, it may be concluded that previous studies on speech recognition using BCI have considered only small numbers of stimuli as well as limited numbers of subjects and different experimental design methods (before this research started in 2015). As a result, it was currently impossible to draw firm conclusions about the possibility of obtaining comparable results across a wide range of stimuli and subjects. The studies presented to date thus are effectively only proof-of-concept for imagined speech recognition and lack complete communication applications.

The experimental designs described in the next three chapters have been proposed based on the research gaps discussed in section 3.6.

Chapter 4

Discriminating between Imagined Speech and Non-speech

As has been discussed in the previous chapters, a brain-computer interface (BCI) can potentially be the only communication option for people who suffer from severe neuromuscular impairments such as locked-in syndrome. Many cognitive tasks have been explored for BCI, ranging from selective attention, motor imagery, and word associations to mental arithmetic (Van Gerven et al., 2009). The use of these modalities for communication can be limiting, however, as they are unintuitive (Van Gerven et al., 2009), they can be limited in the number of classes that may be provided (e.g. only four classes from motor-imagination studies (Lotte et al., 2015)), and/or they require external stimuli (e.g. P300-based BCIs).

In contrast to other instructed-cognitive tasks such as motor imagery, detecting speech imagination is still a new research domain. Numerous questions remain to be answered and identified, including the optimal experimental design, which brain areas are key in capturing the brain activities related to speech, and the effect of phonological and semantic differences between words in the recognition.

Very few studies have focussed on discriminating between imagined speech and non-speech. One such study included a comparison between the imagination of two vowels (/a/ and /u/) by imagined lip movements, using 'no imagination' as a

control state (DaSalla et al., 2009; Yoshimura et al., 2011). Zhao and Rudzicz (2015) investigated three mental states related to speech imagination, actual speech, and stimulus presentation (a word presented on the screen and a sound utterance). In their study, facial expressions and audio signals were combined with EEG signals to improve the classification results. Sereshkeh et al. (2017a,b) used EEG signals recorded from 10-second word repetitions of *"yes"* and *"no"* versus an unconstrained rest time. This setup yielded high classification accuracy compared to previous studies in which two words were classified.

In contrast to the literature, the present work targets a more intuitive imagined-speech procedure that includes imagining words once rather than several times in a fixed time window. The imagination also involves a larger variety of words (11 words and syllables). The words have been selected to be semantically varying to examine the effect of these variations in discriminating words from non-speech tasks. Finally, a low-cost wireless EEG headset was used to record brain signals. These factors all imply a large variation in imagined-speech EEG signals, which makes such speech more difficult to classify.

The focus of this chapter is classification between imagined words versus either relaxation or attention to a visual stimulus. Spatio-spectral and time-domain features were examined for each subject to extract information from the EEG signals. Different intervals were examined for feature extraction. The results are first presented to show how words as a group can be classified from the non-speech class using the proposed features and classification algorithms. The potential of classifying each individual word versus relaxation is then discussed. The remainder of this chapter is structured as follows: section 4.1 reviews the literature on the effect of words' semantics on brain activity. Section 4.2 provides an overview of the design and the procedure used for the experiment presented in this chapter. The section includes information about the participants as well as the EEG device, the stimuli and tasks, the data pre-processing, the feature extraction, and the classification algorithms used in the study. Section 4.3 presents the classification results. Section 4.4 highlights the main conclusions that were

drawn from the experiments results. Finally, section 4.5 summarises the findings from this chapter. Part of the results from this chapter have been presented in a published paper (AlSaleh et al., 2018).

4.1 Effect of word meaning on brain activity

Several studies in the literature have examined the effect of emotional content on the cognitive process. Doerksen and Shimamura (2001), for example, conducted a study to understand the influence of emotional stimuli on source memory. In total 64 words in two sets were presented. One set contained neutral words (e.g. 'chair'), and the other contained emotional words (e.g. 'emergency'). During the study, the participants were asked to read each word silently and to remember the colour in which it had been presented. Generally, the results suggested an enhancement of the source memory for the emotional words, because the participants better remembered the colours in which the emotional words had been typed.

Another study, by Fossati et al. (2003), used fMRI scanning to determine the neural regions involved in the emotional valence of the stimulus. Thirteen lists of ten personality-trait adjectives were constructed from Anderson's list of personality-trait words (Anderson, 1968). This list included 555 personality-trait words rated by 100 subjects based on 'likeableness' as a personality characteristic. The scanning process was conducted three times. First, in the self-referential processing condition, the subjects determined whether they thought each trait described them. Second, in the other referential processing condition, subjects evaluated whether the stimulus represented a generally desirable trait. The third task was letter recognition as a control task. In general the results showed that a widely distributed network of brain areas contributes to emotional processing. Among these regions, the right dorsomedial prefrontal cortex was found to be the main area for self-referential tasks, where subjective, perspective-taking aspects are involved in emotional evaluation.

The objective of (Herbert et al., 2008)'s study was to measure the extent to which emotional connotation influences cortical potentials during reading. To achieve this goal, ERPs were recorded during the reading of high-arousal pleasant and unpleasant adjectives and low-arousal neutral adjectives, presented at rates of 1 Hz and 3 Hz. The words were selected according to the previous independent ratings of 45 subjects on a total of around 500 adjectives. The study demonstrated the effects of emotional word content on a sequence of relatively early posterior negativity (EPN) and late positive potential (LPP; N400) cortical indices during the uninstructed reading of words. In general, the brain initially responds to the emotional significance of a word, regardless of its valence. Schacht and Sommer (2009) followed a similar approach to investigate the changes in event-related brain potentials ERPs related to emotions during visual word processing.

Considering these studies, the hypothesis can be formed that including words with emotional and semantic meanings in the BCI system may improve the speech-recognition system, since different emotions influence the brain patterns in varying ways. To the best of our knowledge, such a study has not yet been undertaken.

4.2 Experiment

4.2.1 Participants

Nine males ranging in age from 18 to 36 participated in this study. Participants with any neurological disorders, a history of brain injury, or a personal or family history of epilepsy, or those who had consumed alcohol or any type of drug in the previous 12 hours, were excluded from the study. The experiment was ethically approved by the Department of Computer Science, University of Sheffield, UK; all participants signed a consent form.

4.2.2 Device

The acquisition of brain signals was performed using an Emotiv EEG neuro-headset. Of the headset's total of 16 channels, 14 channels were used for data recording (AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, and AF4). Two were inactivated to serve as the ground and reference channels.

4.2.3 Stimuli

In this study, 11 words were selected based on variations in their semantic meaning. Section 4.3.4 presents several neuroscience studies that have examined the impact of the emotional implications of words on neural activities as represented by cortical potentials. Syllables were chosen for the 'no semantic' stimuli, which was the approach used in previous studies (DZmura et al., 2009). In the present study, the word stimuli were selected to include emotional words, words with neutral meaning (directions and responses), and syllables, as follows:

- **Syllables**: /ba/ and /ka/.
- **Directions**: *"Left"*, *"Right"*, *"Up"*, *"Down"*.
- **Responses**: *"Yes"*, *"No"*.
- **Emotions**: *"Happy"*, *"Sad"*, *"Help"*.

4.2.4 Task

Before starting to record the EEG signals, the experimental instructions were explained to each participant. These instructions were written out as a script to ensure consistency between all nine participants. The instructions asked the participants to minimise their body movements during the experiment. It was explained that this included hand movements, jaw movements and any other kind of physical movement. Moreover, the participants were instructed to put in emotions for the words that have emotional meaning. The task steps and the stimuli presented are summarised below:

1. **Visual attention (fixation):**

   The symbol, '+', was presented on a screen for one second. The participants were instructed to focus on the symbol.

2. **Relaxation (black screen):**

   In this task, the participants were instructed to relax (be silent) and clear their minds from any type of thinking as much as possible. This task lasted two seconds.

3. **Word presentation:**

   In this task, a word was presented on the screen for two seconds. The presentation of words from this list was performed randomly to avoid the effect of word order.

4. **Word imagination (black screen):**

   Once the screen gets blank, the participants were instructed to immediately imagine the previously presented word for one time. This task lasted two seconds.

A total of 11 imagined speech stimuli were used. The recording was performed as blocks. Six blocks were recorded for each subject. During each block, each word was presented in random order eight times. Hence, a total of 88 fixations and relaxation tasks were conducted for each block (they were presented before and after each word). A total of 48 trials were recorded for each word; and all the stimuli in the experiment consisted of 1584 trials. Figure 4.1 presents the steps of recording one block.

4.2.5 Data pre-processing

High-pass and low-pass zero-phase filters were applied in the range of 1–30 Hz to remove power-line noise and to attenuate noise caused by body movements. For all nine subjects, the F7 and F8 channels were used as ground channels. The AF4 and AF3 channels were removed because they are near the eyes, and most signals recorded from these channels tend to be related to eye blinking and movement (Gupta et al., 2012). Baseline correction was performed to remove the effects that occurred prior to

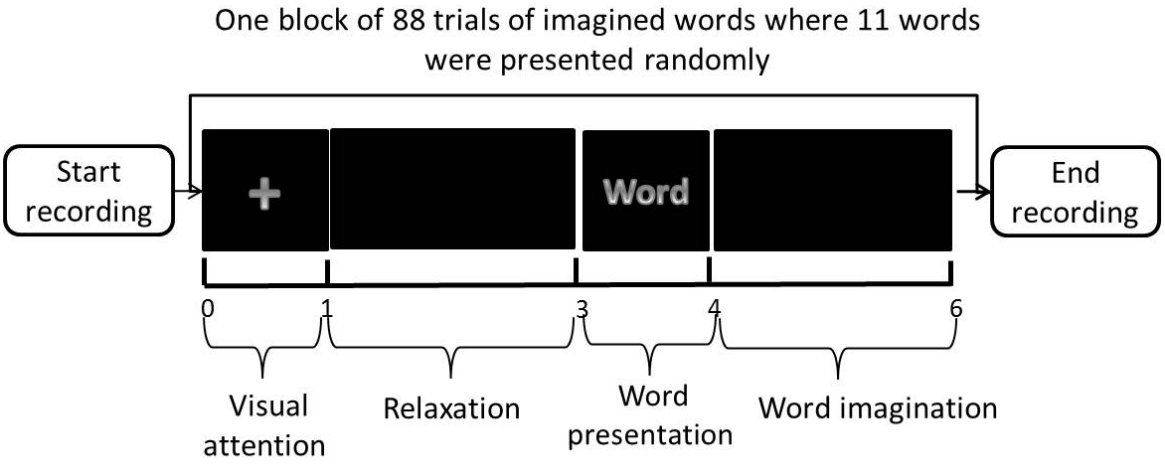

Fig. 4.1 The steps of recording one block (88 trials) of imagined words

the presentation of each stimulus. The baseline may be defined as the time preceding the stimulus. For the present study, the range of -200 ms to 0 ms was removed with respect to the stimulus onset (Woodman, 2010).

4.2.6 Feature extraction

Spatio-spectral and temporal features were investigated during the feature-extraction stage. Time-domain features were extracted by computing four features from each channel: standard deviation (SD), mean, sum of values (SUM), and root mean square (RMS). Spatial features were computed using Ang et al. (2008) filter bank common spatial patterns (FBCSP) algorithm. Both spatio-spectral and temporal features were calculated for three different time intervals after the start of the task: [0–1 s], [0–1.5 s], and [0–2 s]. The following describes how both time-domain features and spatio-spectral features were calculated for each EEG trial.

Time domain features

Time-domain features have been used in several EEG studies in the literature. For example, Kumar et al. (2018) used SD, RMS, SUM, and energy to classify envisioned speech. In the present study, SD, RMS, SUM, and the mean were calculated for the

samples, resulting in four features from each channel. Because 12 channels were used in the current work, a total of 48 time-domain features were generated (12 channels × 4 time-domain features).

Assuming i^{th} EEG trial is presented as $\boldsymbol{X}_i \in \boldsymbol{R}^{n \times c}$ trials set where n is the number of samples, c is the number of channels. Each EEG channel is presented as ch, $\{ch = 1, 2, 3, \cdots, c\}$, each of the time-domain features were computed as:

$$\textbf{Mean } (\boldsymbol{X}_i^{ch}) = \frac{1}{n} \sum_{s=1}^{n} \boldsymbol{X}_{i_s}^{ch}, \tag{4.1}$$

$$\textbf{SD } (\boldsymbol{X}_i^{ch}) = (\frac{1}{n-1} \sum_{s=1}^{n} (\boldsymbol{X}_{i_s}^{ch} - \textbf{mean}(\boldsymbol{X}_i^{ch}))^2)^{\frac{1}{2}}, \tag{4.2}$$

$$\textbf{RMS } (\boldsymbol{X}_i^{ch}) = \frac{1}{n} \sum_{s=1}^{n} \left| \boldsymbol{X}_{i_s}^{ch} \right|^2, \tag{4.3}$$

$$\textbf{SUM } (\boldsymbol{X}_i^{ch}) = \sum_{s=1}^{n} \boldsymbol{X}_{i_s}^{ch}, \tag{4.4}$$

where $s = \{1, 2, \cdots, n\}$, n is $\boldsymbol{X}_i$ length.

Spatio-spectral features

EEG data have poor spatial resolution; therefore, in order to discriminate between the two classes it was necessary to design some spatial filters. Common spatial patterns (CSP) is a well-known spatial filtering algorithm that are based on maximising the variance of one class while minimising it for the other class. Since it has been proposed by Ramoser et al. (2000), several adaptations (Ang et al., 2008) and expansions (Ang et al., 2012) for multi-class classifications were proposed.

The CSP algorithm projects the trial $\boldsymbol{X}$ to the spatially filtered $\boldsymbol{E}$, as:

$$\boldsymbol{E} = \boldsymbol{X}\boldsymbol{W} \tag{4.5}$$

The spatial filter $\boldsymbol{W}$ is a projection matrix that was computed based on simultaneous diagonalization of the covariance matrices from both classes (Ramoser et al., 2000).

As in Ramoser et al. (2000) study, not all the spatial filtered signals were used for extracting features. Instead, only a defined number, m, of the first and last rows of $\boldsymbol{E}$ in equation (4.5) are used for feature extraction. In the present study, m is equal to 2.

Assuming the signals $\boldsymbol{E}_p$ ($p = 1.....2m$) are given, the feature vector $\boldsymbol{F}$ is calculated as:

$$\boldsymbol{F}_p = log(var(\boldsymbol{E}_p)/\sum_{i=1}^{2m} var(\boldsymbol{E}_p)) \tag{4.6}$$

However, CSP may lead to poor classification accuracies if the data is inappropriately filtered with the wrong frequency bands. Ang et al. (2008) proposed that applying a filter bank that filters EEG data into multiple bands can improve the results (filter-bank common spatial patterns (FBCSP)). Seven filters were included in the bank to obtain data ranging between 1 Hz and 30 Hz. This frequency range represents the well-known bands in the literature, and it has been interpreted as delta, theta, alpha, low beta, mid beta, high beta and low gamma. Since the EEG data was filtered using seven frequency bands, and four rows of the CSP filtered signals were considered for each band, the total number of spatio-spectral features was 28.

4.2.7 Classification

The two groups of proposed features (time domain and spatio-spectral) were evaluated separately in different trial lengths of [0–1 s], [0–1.5 s], and [0–2 s]. Using training data, Pearson correlations between features and class labels were calculated for both groups to rank the features. For the classification, 8-fold cross-validation was applied to divide the data into training (80%), development (10%), and testing (10%) data. The development set was used to identify the optimal number of features for every subject that would provide maximum classification accuracy. A support-vector machine (SVM), naive Bayes (NB), random forest (RF), and linear discriminant analysis (LDA) were used.

SVMs depend on the use of a discriminant hyperplane to distinguish between classes. The margins between the classes can be maximised based on hyperplane

selection. This step protects the SVM from over-training sensitivity or the curse of dimensionality (Lotte et al., 2007). In the present study, SVM was applied with linear decision boundaries, which have been shown to be effective in several EEG studies (Lotte et al., 2007; Sereshkeh et al., 2017b).

NB classifiers work based on the assumption that the features related to every data point are strongly or naively independent of each other. NB is among the classifier types that depend on the conditional probabilistic of the Bayes theorem. Each time before a new instance is classified, the probability of each feature is calculated in relation to every class. Thereafter, the instance is assigned to the class with the highest probability (John and Langley, 1995). González-Castañeda et al. (2017) used NB to classify unspoken speech in their work.

RF classifiers create a group of decision trees to vote for the most suitable class. The classifier is created based on a random subset of the training data and randomly chosen features. Each tree then predicts the class as a voting unit. The final decision is based on majority voting. In the current study, 50 trees were used, and the number of variables in each node was $\log_2(Numberoffeatures + 1)$, as suggested in González-Castañeda et al. (2017)'s work. Kumar et al. (2018) also used RF for envisioned speech (object recognition) from EEG signals.

LDA classifiers are similar to SVMs in the use of a hyperplane to separate the classes. LDA works based on the assumption that the data is normally distributed with an identical covariance matrix for both classes (Lotte et al., 2007). Separations between the two classes are achieved by finding which projections reduce in-class variance and increase between-class means. In the case of multi-class classification, several hyperplanes are used. LDA is simple, has relatively low computational requirements, and has been successfully applied in several EEG studies (Lotte et al., 2007). LDA is sensitive to the dimensionality of the classified data in relation to the proposed features, however. One common problem in domains with small data sizes is known as the singularity of the within-class scattering matrix, which is caused by high dimensionality (Huang et al., 2002).

4.3 Results and discussion

4.3.1 Classifying group of words vs non-speech

EEG trials related to word imagination were labelled as one group (class) and classified against the non-speech tasks (either visual attention or relaxation). A total of 528 speech trials were used, and the same number was used for each of the non-speech tasks. Visual attention was related to two stimuli: the '+' symbol and word presentation. Because it is two classes classification, the random baseline is 50%.

As shown in Tables 4.1 and 4.2, on average the classification accuracies between visual attention and imagined speech were better than those between imagined speech and relaxation across all classifiers at all time intervals when FBCSP features were used. This result makes sense, as visual attention provokes visual processing in the brain that is absent from speech imagery and relaxation. Pairwise t-testing, however, showed that these results were not statistically significant in all cases, as shown in Table 4.3. In both classifications of relaxing and visual attention from groups of words, the maximum average classification accuracy for each classifier was achieved in the time interval [0–1 s].

Among the four classifiers, RF provided the maximum average accuracy in the time interval [0–1 s]: 63.14% and 67.15% for relaxing and visual stimuli, respectively. However, applying paired t-test shows that, the classification accuracy using the time interval [0–1 s] was not statistically different from using longer time intervals in classifying the group of imagined words versus the relaxation task. For classifying visual attention from groups of words, the time interval [0–1 s] significantly outperformed the [0–2 s] interval for SVM, NB, and LDA classifiers.

Table 4.1 Average 8-fold cross-validation results (%) of classifying relaxing (non-speech) and all imagined words using Filter-bank CSP features

Subject	SVM			NB			RF			LDA		
	1 s	1.5 s	2 s	1 s	1.5 s	2 s	1 s	1.5 s	2 s	1 s	1.5 s	2 s
1	69.32	69.89	70.45	69.41	71.4	70.74	70.55	69.32	72.06	69.89	70.08	71.78
2	67.8	67.52	62.12	67.71	66.38	62.88	69.41	68.37	62.41	67.61	66.38	62.22
3	53.64	57.5	56.59	53.98	57.84	58.41	60.00	58.64	58.41	54.89	58.64	56.82
4	57.01	58.24	57.58	56.06	57.58	57.95	54.83	56.34	56.63	55.40	57.48	57.95
5	60.80	62.31	60.32	61.27	61.93	61.08	62.22	61.27	59.85	61.27	62.12	59.00
6	66.00	65.25	66.38	66.57	65.81	66.19	62.78	65.63	66.00	66.57	64.77	66.67
7	62.39	59.20	60.68	58.75	57.84	59.43	64.89	65.57	65.45	61.59	59.32	58.86
8	62.97	60.89	59.19	62.50	59.47	59.38	61.74	59.09	56.63	63.07	60.23	59.85
9	58.64	57.50	59.09	58.52	58.07	58.98	61.82	60.91	62.95	57.95	57.39	57.73
AVE	62.06	62.03	61.38	61.64	61.81	61.67	63.14	62.79	62.27	62.03	61.82	61.21
SD	4.85	4.31	4.17	5.06	4.69	4.03	4.48	4.32	4.77	5.02	4.15	4.68

Table 4.2 Average 8-fold cross-validation results (%) of classifying visual attention (non-speech) and all imagined words using Filter-bank CSP features

Subject	SVM			NB			RF			LDA		
	1 s	1.5 s	2 s	1 s	1.5 s	2 s	1 s	1.5 s	2 s	1 s	1.5 s	2 s
1	66.86	66.57	64.87	65.34	65.44	64.3	66.57	67.23	62.12	67.42	65.72	65.06
2	74.81	76.80	73.48	75.09	76.23	72.54	73.11	75.85	71.02	75.19	78.13	72.44
3	71.25	66.82	64.66	71.25	66.48	65.34	70.57	67.39	69.32	70.80	66.59	65.34
4	57.67	56.16	54.64	56.91	57.10	55.97	58.52	57.39	57.67	59.09	58.43	55.49
5	73.39	65.72	58.90	71.31	62.69	58.24	72.25	62.59	57.29	71.59	62.31	57.67
6	68.75	65.34	64.77	70.36	65.81	64.49	68.47	65.72	62.31	69.13	64.68	63.92
7	61.70	60.80	58.98	60.11	59.89	60.45	66.02	67.73	66.48	60.34	60.23	60.45
8	60.34	60.23	60.45	66.00	68.84	67.52	66.00	67.61	68.37	65.25	70.27	66.86
9	63.52	53.3	50.91	62.39	58.07	48.07	62.84	59.55	56.59	62.5	53.75	49.55
AVE	66.48	63.53	61.30	66.53	64.51	61.88	67.15	65.67	63.46	66.81	64.46	61.86
SD	5.70	6.51	6.22	5.63	5.61	6.76	4.36	5.08	5.23	5.13	6.67	6.46

Table 4.3 Pairwise t-test for each classifier to compare between the classification accuracies in different time intervals of classifying group of words versus relaxing and versus attention to visual stimuli;✓ means significant with p value, $\times$ means not significant

	Time interval		
Classifier	[0-1 s]	[0-1.5 s]	[0-2 s]
SVM	$\times$	$\times$	$\times$
NB	✓(0.04)	$\times$	$\times$
RF	✓(0.02)	$\times$	$\times$
LDA	✓(0.03)	$\times$	$\times$

4.3.2 Spatio-spectral features vs time domain features to classify imagined words vs relaxation

Time-domain features were then computed to classify between imagined words versus relaxation (see Table 4.4). The average accuracy across subjects in both classifiers and all time intervals was less than the classification of FBCSP features. Pairwise t-testing showed that this outperformance of FBCSP was statistically significant only with the SVM classifier in the time intervals [0–1 s] and [0–1.5 s], where the p values were equal to 0.03 and 0.01, respectively. The results were different between subjects, however. For example, for S1 and S3, the time-domain features yielded less accurate results, whereas for S2 and S4, these features yielded better results. This finding suggests that applying a feature-selection method to select between time-domain and spatio-spectral features would further enhance the results.

Table 4.4 Average 8-fold classification accuracy (%) between relaxing (non-speech) and all imagined words using time domain features

Subject	SVM			NB			RF			LDA		
	1 s	1.5 s	2 s	1 s	1.5 s	2 s	1 s	1.5 s	2 s	1 s	1.5 s	2 s
1	48.11	49.34	48.77	51.70	52.75	51.33	51.14	51.61	52.27	55.68	54.26	52.94
2	63.45	65.53	64.11	69.60	67.80	64.20	71.69	67.71	66.95	69.70	67.23	65.15
3	52.05	47.84	52.84	54.09	53.52	54.66	57.05	57.84	59.89	52.27	52.84	53.3
4	58.81	56.72	57.20	60.04	60.98	59.56	63.73	66.57	66.00	62.69	64.87	66.29
5	54.55	53.41	54.17	55.40	55.78	56.53	59.00	61.36	56.72	54.26	55.68	54.55
6	54.26	55.68	54.55	62.97	62.69	61.74	65.25	64.87	63.45	63.35	62.41	62.69
7	50.57	51.59	52.61	55.45	53.64	52.84	66.70	67.05	65.57	56.82	55.23	55.57
8	67.23	65.91	66.86	67.14	65.06	65.34	65.34	64.87	63.83	66.86	65.15	66.67
9	51.52	49.34	48.86	52.84	51.8	51.04	50.76	52.08	50.76	53.41	52.08	51.61
AVE	55.62	55.04	55.55	58.80	58.22	57.47	61.18	61.55	60.60	59.45	58.86	58.75
SD	5.97	6.35	5.90	6.11	5.64	5.15	6.77	5.94	5.72	5.98	5.63	5.95

4.3.3 Combining spatio-spectral features and time domain features to classify imagined words vs relaxation

As what has been described in section 4.3.2, the best feature to classify between imagined words vs relaxation is different for each subject. To examine the gain from combining both features, Table 4.5 lists the average classification accuracies resulted

from the feature fusion. In comparison to the results for each feature separately (Tables 4.1 and 4.4), the average classification accuracies across all subjects are close to the results of using FBCSP feature only. Although for each subject the results are close to the results gained from the best feature for that subject. For example, for subject-1 the results are close to using FBCSP and for subject-8 the results are close to using time-domain features.

Table 4.5 Average 8-fold classification accuracy (%) between relaxing (non-speech) and all imagined words using both time domain features and FBCSP features

Subject	SVM			NB			RF			LDA		
	1 s	1.5 s	2 s	1 s	1.5 s	2 s	1 s	1.5 s	2 s	1 s	1.5 s	2 s
1	69.03	72.25	69.03	69.13	71.40	69.70	69.13	70.17	67.80	70.17	70.93	68.66
2	71.69	66.29	61.84	69.60	62.78	59.09	71.12	69.32	62.78	70.83	66.67	63.26
3	58.86	57.16	53.18	60.23	58.07	56.48	60.00	58.52	61.59	58.75	56.82	54.43
4	57.29	56.06	56.91	55.02	55.78	57.39	59.47	57.2	60.51	55.21	54.83	56.63
5	62.50	62.31	62.12	62.88	63.07	59.85	63.73	60.98	58.05	62.50	63.73	62.03
6	65.63	63.83	66.19	68.37	68.09	67.71	65.91	64.77	64.68	66.67	64.77	65.81
7	56.59	58.98	56.93	56.59	56.82	58.86	64.43	65.57	67.27	57.50	57.84	59.55
8	68.28	67.23	65.53	67.05	65.34	66.86	67.33	63.26	65.25	67.52	65.91	66.95
9	50.76	53.41	53.6	49.43	55.21	53.88	50.76	53.69	55.30	50.28	55.40	55.02
AVE	62.29	61.95	60.59	62.03	61.84	61.09	63.54	62.61	62.58	62.17	61.88	61.37
SD	6.92	6.08	5.72	7.2	5.76	5.58	6.14	5.53	4.18	7.17	5.77	5.27

Table 4.6 Number of words that provide above (58%) classification accuracy against relaxation using filter-bank CSP features

Subject	SVM			NB			RF			LDA		
	1 s	1.5 s	2 s	1 s	1.5 s	2 s	1 s	1.5 s	2 s	1 s	1.5 s	2 s
1	5	6	11	6	7	9	7	7	8	2	7	7
2	6	4	4	6	5	5	5	4	6	6	6	3
3	0	2	3	1	2	3	0	2	3	0	0	3
4	0	5	1	1	2	1	1	1	0	1	4	0
5	5	3	2	5	4	4	5	2	3	5	3	3
6	5	5	3	5	5	4	4	4	4	5	3	3
7	4	3	3	2	2	3	1	2	1	4	3	2
8	3	4	4	5	4	4	5	4	4	4	4	4
9	0	2	3	2	4	4	1	0	0	1	3	3
AVE	3	4	4	4	4	4	3	3	3	3	3	3

Table 4.7 Number of words that provide above (58%) classification accuracy against relaxation using time-domain features

Subject	SVM			NB			RF			LDA		
	1 s	1.5 s	2 s	1 s	1.5 s	2 s	1 s	1.5 s	2 s	1 s	1.5 s	2 s
1	1	1	1	0	3	2	5	6	6	5	5	7
2	6	4	7	1	0	1	4	1	3	9	8	8
3	2	1	2	1	1	4	0	2	3	2	2	2
4	3	2	2	4	6	4	4	7	6	5	5	6
5	0	2	2	3	2	1	1	2	4	2	3	3
6	8	5	3	8	7	7	6	8	4	6	8	5
7	3	2	3	5	4	4	3	3	4	5	5	3
8	9	9	11	11	10	11	10	10	11	11	10	10
9	1	0	0	0	2	4	1	1	2	3	2	2
AVE	4	3	4	4	4	4	4	4	5	5	5	5

4.3.4 Classification of individual words versus relaxation

Each individual word was imagined in 48 trials during the experiment. During the classification of each word versus relaxation, the 48 trials of the imagined words were compared with the 48 trials of relaxation that occurred before the same word. Using the binomial cumulative distribution (Combrisson and Jerbi, 2015), the upper limits of 95% confidence interval of chance was $\approx$ 58%. Tables 4.6 and 4.7 show the number of words that provide above 58% 8-fold average classification accuracy using FBCSP and time-domain features using SVM, NB, RF, and LDA in different trial lengths for each word versus relaxation time. The results were very encouraging, as the study used only a small number of training trials, a low-cost EEG device, and single imagination repetition. One interesting finding is that, in comparison with the results shown in Tables 4.1 and 4.4, the classification of all words as one group was found to help in identifying the best features for each subject. For example, for S1, the best feature set was found to be FBCSP, which was supported by the number of classified words. Otherwise, for S4 and S2, time-domain features may be concluded to provide the best number of classified words. These findings make the classification of groups of words

versus relaxation time an important step in measuring the effectiveness of the selected feature.

In Table 4.6, the maximum number of words classified against relaxing time in above than 58% using FBCSP is 4. Figure 4.2 presents the distribution of average classification accuracy for each word versus relaxing for all subjects using FBCSP in time interval [0-2 s] using SVM classifier. For time-domain features (Table 4.7), the maximum number of words classified against relaxing time in above than 58% is 5. Figure 4.3 presents the distribution of average classification accuracy for each word versus relaxing using LDA classifier in time interval [0-2 s] for all subjects.

The analysis of word classification also leads to the conclusion that word semantics have no effect on classification accuracy, although the subjects were instructed before the experiment to add emotions during their imagination of the word. This result may be justified from two perspectives: task difficulty and experiment design. The recognition of acted emotions from audio (i.e. audio speech) is often considered to be a difficult task (Bachorowski, 1999; Banse and Scherer, 1996). Usually this type of experiment requires the hiring of an actor to perform it. From the experimental-design point of view, giving the participants two seconds may not have been sufficient for them to produce this emotional speech, taking into account that the words were presented randomly, and the emotional level in their meanings varied. For future improvement of the recognition of semantically varying words, the recognition of emotions from EEG signals can be included to enhance the recognition. Several studies about the recognition of emotions from EEG signals have been conducted in the literature as in (Yohanes et al., 2012).

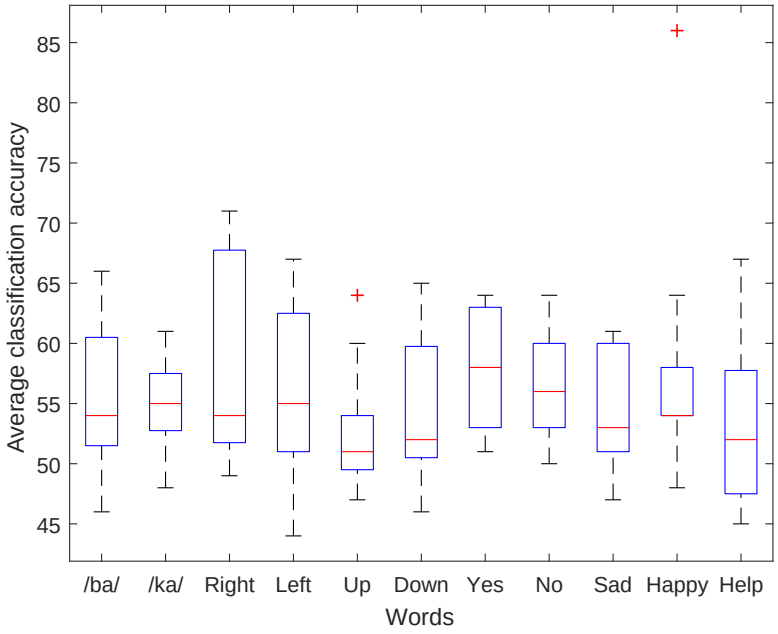

Fig. 4.2 Average 8 fold classification accuracy between relaxing and each word using FBCSP features, SVM classifier in time interval [0-2 s] for all 9 subjects

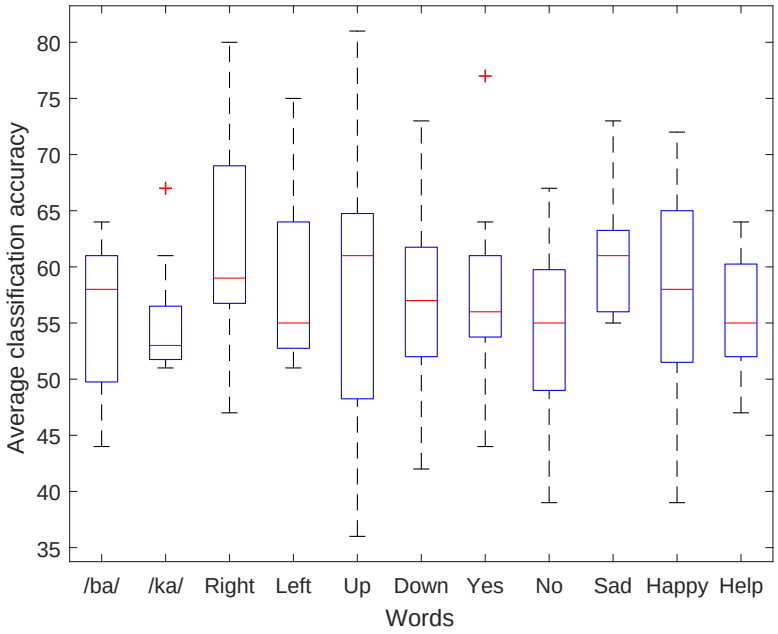

Fig. 4.3 Average 8 fold classification accuracy between relaxing and each word using Time domain features, LDA classifier in time interval [0-2 s] for all 9 subjects

4.4 Conclusions

The study presented in this chapter has investigated the possibility of discriminating between imagined speech and two types of non-speech tasks related to either attention to visual stimulus or relaxation. The results were discussed from two main perspectives: classifying non-speech tasks versus group of words and versus each word individually.

For classifying the non-speech tasks versus group of words, the maximum average classification accuracy was 67.15%. This result for two class classification is not considered high as the random baseline is 50% for two classes classification problems. This could be justified as a consequence of the difficulty of the task design. This experiment has examined semantically varying words (11 words and syllable) with a single imagination in each trial during two seconds. The EEG data was also recorded from a small number of electrodes using a portable and inexpensive EEG device. Thus, the experiment design was closer to what we want to achieve in the future as communication tool for locked-in patients. However, this design makes the EEG classification more challenging due to a higher level of noise and variations in EEG signals.

For the classification of each word versus relaxing, the experiment was designed based on the studies discussed in Section 4.1 that showed the potential effect of words semantics on distinguishing EEG patterns. Based on that, four different categories of semantically varying words have been selected (see section 4.2.3). The classification accuracies of each word versus relaxation did not help in drawing any conclusions about the differences between the recognition of the words. As has been discussed in Section 4.3.4, the evoking of emotions is a challenging task and several design issues should be taking in to account.

4.5 Summary

This study presented a first step in understanding how imagined speech can be recognised from another tasks using only EEG data. The imagined speech was compared to attention to visual stimuli and to relaxing. In comparison to previous studies, this study stimuli were semantically varied with only single imagination in each trial. Spectral and temporal features, with and without common spatial filtering, were used for classifying every imagined word (and for a group of words) against the non-speech tasks. Both features were extracted from three different time intervals: [0-1 s], [0-1.5 s], and [0-2 s]. The results vary across subjects and according to different types of tasks. In this study, no effect of words semantics were found in the classification of each word versus relaxing. The following chapter will investigate the classification between different imagined words and the effect of temporal parameters in improving this classification.

Chapter 5

Examining Temporal Issues Related to Imagined Speech Recognition

Motor imagination is one of the neural activities in BCI that is well examined in the literature as a potential technology to help paralyzed people to interact with the external world. Commonly, these studies examine the classification between the imagination of the movement of the right hand, left hand, tongue and feet. In motor imagination experiments, the participants are asked to perform the motor imagination task continuously for a specific amount of time. For example, in the most popular dataset for motor imagination, the length of imagining each body movement was 2.75 seconds (Brunner et al., 2008). In general, motor imagination lends itself well to being continuously reproduced as the patterns can be consistently repeated.

For speech imagination, several studies use EEG to capture imagination of pronouncing words (González-Castañeda et al., 2017; Porbadnigk et al., 2009; Suppes et al., 1997), syllables (DZmura et al., 2009) and vowels (Yoshimura et al., 2011). In comparison with the motor task, the speech task is discrete and short. The normal speech rate is 120-180 words per minute, about 0.33-0.5 seconds for every word (Miller et al., 1976). This rate is around five times larger than that of the motor imagination

task described in (Brunner et al., 2008). As a result, capturing EEG patterns related to speech events is challenging. The nature of the speech task influences the design of unspoken speech studies to get consistent and sufficiently long patterns.

In the literature there are several differences between unspoken speech studies. These differences are mainly related to the length and repetition of speech task. This chapter focusses on the recognition of unspoken words using block recording mode and how the temporal experimental parameters can improve the recognition (the differences between recording modes were discussed in Section 3.6).

All of these previous studies are not consistent from two experiment design perspectives:(a) the number of trials each word should be imagined (training size), and (b) the length of the imagination. The first perspective was examined partially in (Porbadnigk et al., 2009) for the recognition of five words. The recording for every word was performed in four modes: long blocks (20 repetitions), short blocks (5 repetitions $\times$ 4 blocks) or a single pronunciation of ordered or randomised words for a total of 20 trials for each word. The results showed that only the long-block recording resulted in an accuracy of 45% for 5 words. Furthermore, a cross-session examination was conducted for two participants. The results show a chance level when the training was performed in one-session blocks and the testing in another session blocks. In this work (Porbadnigk et al., 2009), the researchers justified that the temporal correlation between the trials in the long blocks makes the recognition rate higher than short blocks or individual words imagination.

This chapter focuses on EEG based unspoken words recognition using block recording to address the following questions:

1. How does the choice of word separation technique affect the classification accuracy?

2. What is the relation between the number of repetitions (training size) and the classification accuracy?

3. How does the repetitions order affect the classification accuracy?

4. How does the determination of the exact time of speech imagination change the classification accuracy?

The answers to these questions are important for improving recognition of unspoken speech as the EEG data is known to vary between/within sessions and the recording of a large amount of training is impractical. Moreover, long calibration time and long recording sessions might affect the quality of the data due to fatigue.

To answer the above listed questions, EEG data during the imagination of five words were collected. The recording was divided into two parts. In the first part, EEG trials were separated using mouse clicks, where the subject had to preform one click before and after each trial. In the second part, the subject was given a fixed time to preform the imagination task in each trial (more details are described in section 5.1).

This chapter is structured as follows. The experiment design including (participants, EEG device, stimuli and task, and experiment procedure) are described in section 5.1. Section 5.2 explains the data analysis. Section 5.3 presents and discusses the results. Section 5.4 highlights the main conclusions from the experiments results. Finally, Section 5.5 summarises the findings from this experiment.

5.1 Experiment design

5.1.1 Participants

The study was approved by the ethical committee of Department of Computer Science, University of Sheffield, UK. All the participants have signed the informed consent form. Ten males participated, and they were in the age range of 18-36 (Mean=22, SD=4.6). Six of them were native speakers, and four had studied English for an average of ten years. All the participants disclosed that they were not suffering from any neurological, psychological or heart problems and had not consumed any drugs or alcohol in the 12 hours before the session time.

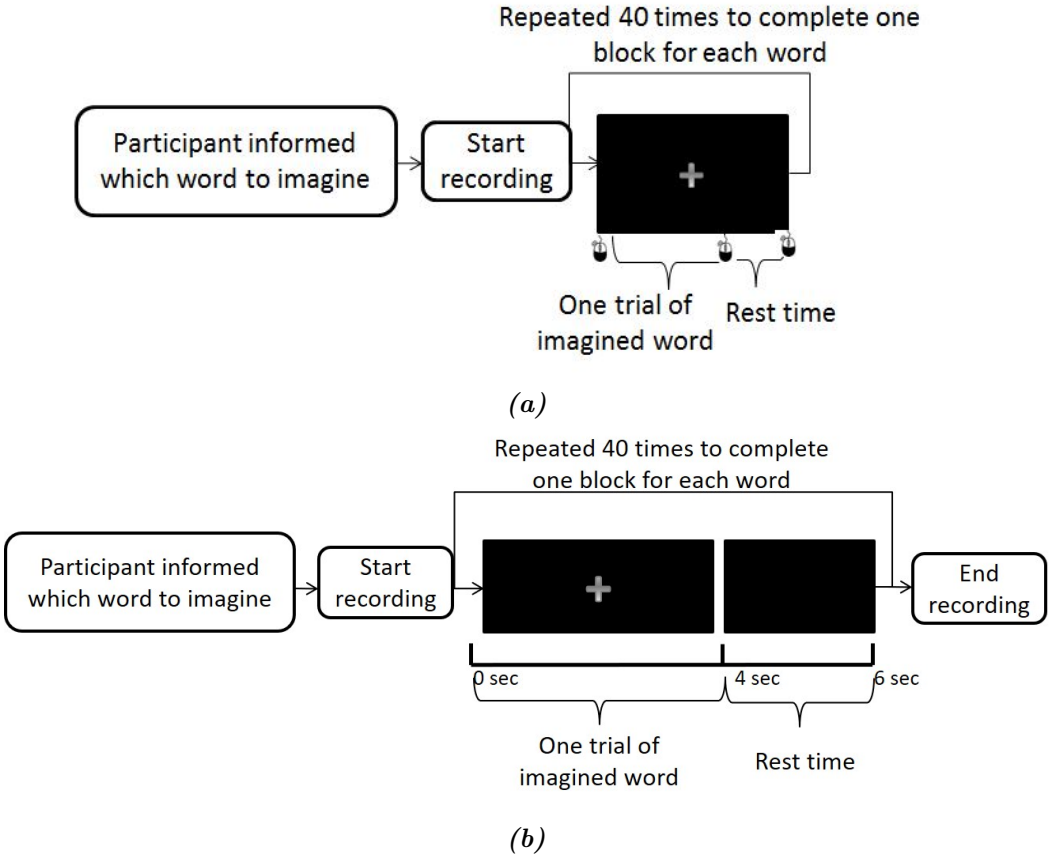

Fig. 5.1 The difference between mouse clicks trials separation (a) and fixed time window trials separation (b).

5.1.2 EEG Device

The Emotiv Epoc headset was used to record EEG data at a sampling rate of 128 Hz. This headset is a wireless device that consists of 14 channels. Based on the 10-20 system (Jasper, 1958b), these channels are AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8 and AF4.

5.1.3 Stimuli and task

The following five words were chosen: *"left"*, *"right"*, *"up"*, *"down"* and *"select"*. These words could be used to control mouse cursor. In previous studies the recognition of these words was examined in (Antonio et al., 2012; González-Castañeda et al., 2017) for the Spanish language.

The participants were asked to imagine the pronunciation of each word for a total of 100 trials (repetitions) during the recording session. The participants were instructed not to move any muscles or blink their eyes during the imagination period (trial). The recording was divided into two parts on the basis of how the trials were separated:

1. Mouse clicks: sixty trials (divided into two block of 40 and 20) were collected for every word. The participant made one mouse click immediately before and after each trial (i.e the word imagination period). During the recording, the time between the end of one trial and the start of the next was decided by the participant and could be used as the rest time for the participant.

2. Specified time frame: forty trials for every word were collected as a block. The participants were given four seconds to imagine the pronunciation for each word followed by two seconds as the rest time between trials.

5.1.4 Procedure

Five participants started with the mouse click method, and five started with the time frame method. The purpose was to remove the effect of time and fatigue on the recognition rate. Below, the steps are explained:

Mouse clicks trials separation

- The participant sat in front of a black screen which had a grey "+" symbol on it, and was informed which word he had to pronounce.

- When the recording started, the program counted 40 trials of that word based on the number of clicks.

- The trial started when the participant made the first click, performed the imagination and then made the second click.

- After recording, one block of 40 trials for every word in the following order: *"left"*, *"right"*, *"up"*, *"down"* and *"select"*. Another block for every word, including 20 trials, was recorded. However, the order of words was changed to the following to remove the effect of word order: *"up"*, *"down"*, *"select"*,*"right"* and *"left"*.

Time frame trials separation

- The trial started when "+" appeared on the screen for four seconds. The participant had to imagine the pronunciation of the identified word during the four seconds period. When the "+" sign disappeared, it meant a two-second rest time for the participant.The order of the words was *"left"*, *"right"*, *"up"*, *"down"* and *"select"*.

5.2 Data analysis

5.2.1 Pre-processing

The data was filtered using a Butterworth (0.5-43 Hz) zero-phased band-passed filter to remove any powerline noise, and reduce the effect of electrooculography (EOG) or electromyography (EMG) artefacts. After that the trials were extracted from the available channels. For all subjects, channels F7 and F8 were used as ground, whereas AF4 and AF3 were excluded as they mostly recorded eye movements and blinks. For the mouse click trials separated data, the trial was taken to be the samples between two clicks. For the fixed time frame data, the trial was taken to be the samples during displaying "+". For every trial, baseline correction was performed by subtracting the average EEG for 200 ms before the trial. This was to ensure that there is no overlap between the EEG signals of interest and the EEG signals that happened before (Woodman, 2010).

5.2.2 Feature extraction

Discrete Wavelet Transform (DWT) has been applied in several EEG studies. For example, epileptic seizure detection (Subasi, 2007), unspoken speech recognition (Antonio et al., 2012; González-Castañeda et al., 2017), emotion recognition (Angrisani et al., 1998; Yohanes et al., 2012). DWT decomposes the signal into detailed and approximation coefficients by analysing the signal into different frequency bands. This is performed by consecutive high-pass and low-pass filters which are based on a selected mother wavelet. In EEG studies, Daubechies2 (db2) or Daubechies4 (db4) have been used as the mother wavelet.

In this study (db4) was used with five decomposition levels as this was proposed in (Sereshkeh et al., 2017b) and (Sereshkeh et al., 2017a) for classifying between two words ("yes" and "no"). However, in this work numbers of resulting wavelet coefficients is different because the participants can perform the imagination in different time lengths. To make the number of features identical for all trials, in (González-Castañeda et al., 2017; Guo et al., 2009) it has been proposed to calculate the Relative Wavelet Energy (RWE) for all the detailed coefficients and the approximation coefficient to equalize the number of features (see section 6.4.2). However, the calculation of energy includes summation of DWT coefficients which reduces the effectiveness of DWT because it removes the temporal information included in the coefficients (Yohanes et al., 2012). Therefore, statistics on the DWT coefficients were applied as proposed in (Sereshkeh et al., 2017b) and (Sereshkeh et al., 2017a). More specifically, standard deviation (SD) and root mean square (RMS) of DWT from every channel were calculated. Moreover, the pilot analysis showed that compared to RWE these statistics on DWT lead to better classification results.

As there were 12 channels involved, with 6 DWT decomposition levels (five detailed coefficients and one approximation coefficient) from the DWT, the total number of features is 144 (12 EEG channels $\times$ 6 decomposition levels $\times$ 2 features i.e. SD and RMS). In addition, for the mouse click separated trials the number of samples between the start and the end click was counted as the imagination length feature.

5.2.3 Classification

Four classifiers were trained: support vector machine (SVM), naïve bayes (NB), random forest (RF), and linear discriminant analysis (LDA) (see Section 4.2.7 for more details about the classifiers). In this study the classification models were subject dependent and 10-fold cross validation was used to evaluate them. However, there was a difference in how training and testing sets were selected in each part as discussed in the following sections.

5.3 Results and discussion

5.3.1 Classifying between the five imagined words

As it has been explained in the experiment procedure, the participants pronounced the words in blocks where each block represent specific word trials. For every word there are two methods to separate the trials: mouse click and 4 seconds fixed time frame. Table 5.1 presents the average 10-fold classification accuracy between the five words for the two separation methods using four different classifiers. For every word in each method, 35 trials were used for training and 5 trials for testing, all from the same block. Interestingly, for all the classifiers using a fixed time frame gives higher average classification accuracy. The maximum accuracy is 98.47% using RF for subject 4 and the lowest accuracy was 40.21% using SVM for subject 10. However, for subject 1 and 10 in some cases the mouse click separated data outperform the fixed time frame separated data.

Table 5.2 shows that the differences between the classification accuracies of the fixed time frame separated data and the mouse click separated is statistically significant for all the classifiers except LDA. This significant out-performance of the fixed time separation approach can be explained from two perspectives. First, the mouse click separated data includes some activities related to the intention to click and the click itself. In addition, the compared fixed time frame is 4 seconds which is relatively long

Table 5.1 10-folds average classification accuracy to classify between five words for mouse click separated data and fixed time frame separated data; the best result for every subject is in bold

	Mouse Click				Fixed Time Frame			
Subject	SVM	NB	RF	LDA	SVM	NB	RF	LDA
S1	68.83	73.07	**87.22**	58.71	61.29	77.34	86.42	49.71
S2	41.78	52.92	57.11	45.53	68.84	82.89	**84.42**	67.34
S3	50.32	64.06	69.82	54.97	60.79	72.37	**88.95**	58.29
S4	61.35	78.86	79.3	53.95	68.26	90.97	**98.47**	74.31
S5	85.2	90.52	92.11	74.59	87.5	85.58	**92.46**	40.33
S6	37.05	44.42	54.65	33.83	55.37	76.89	**80.39**	51.84
S7	67.28	52.95	70.38	51.26	87.42	82.95	**93.95**	76.47
S8	48.60	54.47	60.88	46.08	68.37	66.87	**83.47**	54.87
S9	50.23	67.22	71.96	46.02	83.94	95.00	**97.00**	83.95
S10	49.77	67.16	73.07	56.64	40.18	59.76	**73.79**	40.21
Average	56.04	64.57	71.65	52.16	68.20	79.06	87.93	59.73

in comparison to the maximum time every subject needed to do the imagination. More discussion about the effect of time frame length is given in sections 5.3.2 and 5.3.4.

RF outperforms all classifiers in both mouse click and fixed time frame separated data. Calculating the confusion matrix for RF classifier for both mouse click and fixed time separated data (see Table 5.3 and Table 5.4) shows that word "Left" has the best classification accuracy. Words "Right" and "Up" have very close classification accuracy. Although, this conclusion is made from the average confusion matrix where this might be different for each subject.

These five words have been selected in light of recent study (González-Castañeda et al., 2017) as they can be used in future applications of controlling tasks. However, there is no clear evidence from the literature about the relation between these words classification and motor imagination related brain areas (Qureshi et al., 2018). In the next chapter, the classification of these words is examined using the proposed DTW.

Table 5.2 Pairwise T-test for each classifier to compare between the classification accuracies obtained by the mouse click trials separation data and the fixed time trials separation data;✓ means significant with p value, $\times$ means not significant

Classifier	T-test
SVM	✓(0.02)
NB	✓(0.01)
RF	✓(0.001)
LDA	$\times$

Table 5.3 Confusion matrix (%) for classifying the five imagined words using RF classifier for mouse click separated data

Word	Left	Right	Up	Down	Select
Left	80.00	9.25	5.00	4.00	1.75
Right	10.75	68.00	10.75	5.50	5.00
Up	6.75	10.50	68.50	7.50	6.75
Down	6.50	10.00	9.50	63.50	10.50
Select	1.50	6.00	4.50	9.00	79.00

Table 5.4 Confusion matrix (%) for classifying the five imagined words using RF classifier for fixed time frame separated data

Word	Left	Right	Up	Down	Select
Left	92.00	5.50	1.25	0.5	0.75
Right	3.25	89.75	3.5	1.5	2.00
Up	1.75	4.75	89.5	2.25	1.75
Down	0.75	2.75	4.25	84.5	7.75
Select	1.11	3.61	2.2	9.72	83.33

5.3.2 Effect of training size

To examine the effect of training size on the classification accuracy, 10-fold cross validation was performed for the mouse click separated trials data as in each fold 5 trials per word were used for testing while the training size was varied between 5, 10, 15, 20, 25, 30, and 35 trials per word. The four classifiers were trained using variable sized data where the trials of each word came from the same block.

Figure 5.2 shows the average cross-validation classification accuracies of the four classifiers across different size of training set. As can be seen, the highest improvement for SVM, NB, and RF was obtained by increasing the number of training trials from 5 to 10 per class. Thereafter, for the SVM classifier the improvement is continued and the maximum accuracy is obtained by using all 35 trials per class in training. For NB and RF, the maximum accuracy is nearly achieved by using 30 trials per class for training. Interestingly, in NB and RF, the improvement in the average accuracy is less the 2.00% after using 20 trials per class for training.

LDA behaved differently compared to the other classifiers where the maximum accuracy was achieved with less training data and the accuracy degraded until having 30 trials in training. Thereafter, the average accuracy increased with 35 training trials from every class. This is because of the problem of singularity of the within-class scatter matrix that appears due to few training data (Huang et al., 2002; Markopoulos, 2017). As a result, the reliable results of LDA starts with having 35 trials in training as the number of training trials (175) becomes more than the number of features (144).

For the fixed time frame separated data, the improvement in accuracy was evaluated from two perspectives: training size, and frame length. Similar to the mouse click separated data, the training size was varied, however, each analysis was repeated using different imagination time frames as the trial length (i.e. 0.5, 1, 1.5, 2, 2.5, 3, 3.5, and 4 seconds immediately started from the beginning of the imagination). In figure 5.3, the behaviour of each classifier is presented. As expected, for SVM, NB and RF the average accuracy increases with the increase of training size regardless of the length of the time frame. Interestingly, increasing the length of the time frame also leads to

an increase in the accuracy, although the results of the 3.5 and 4 seconds time frames are very closed (0.3 % average difference). The relation between the increase in the time frame and the improvement in the classification accuracy can be justified as a longer time frame could improve the estimation of DWT. This might be similar to the concept of wavelet zero-padding (Pardey et al., 1996) as baseline correction was performed and the participants were instructed to perform the imagination at the beginning of the time frame and have clear mind after that. As a result, the end part of the time frame is most-likely similar to adding zeros to the end of the time frame. Further investigation is needed to prove this hypothesis. Similar trend is observed for all the classifiers except LDA, perhaps because LDA is more affected by training size as previously explained for the mouse click separated trials data.

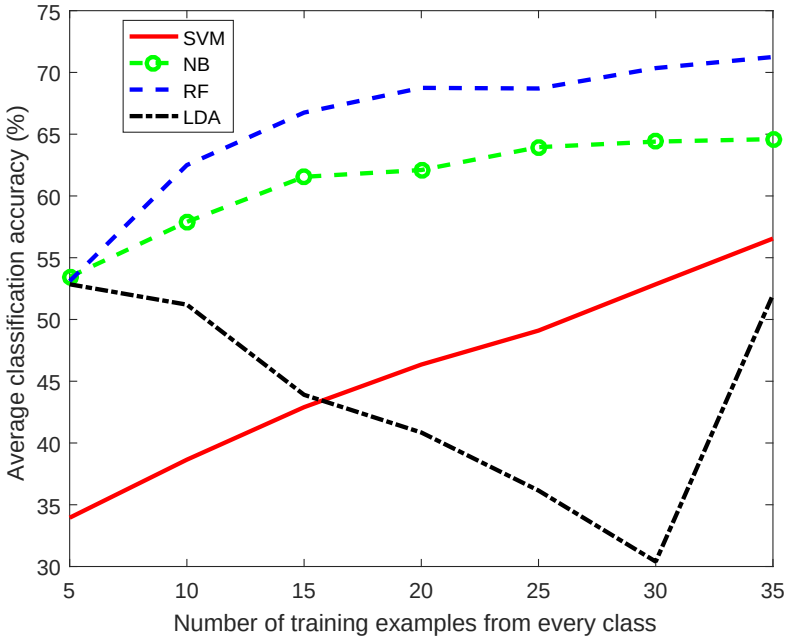

Fig. 5.2 Average 10-fold classification accuracy (%) using different training sizes for MC data using different classifiers.

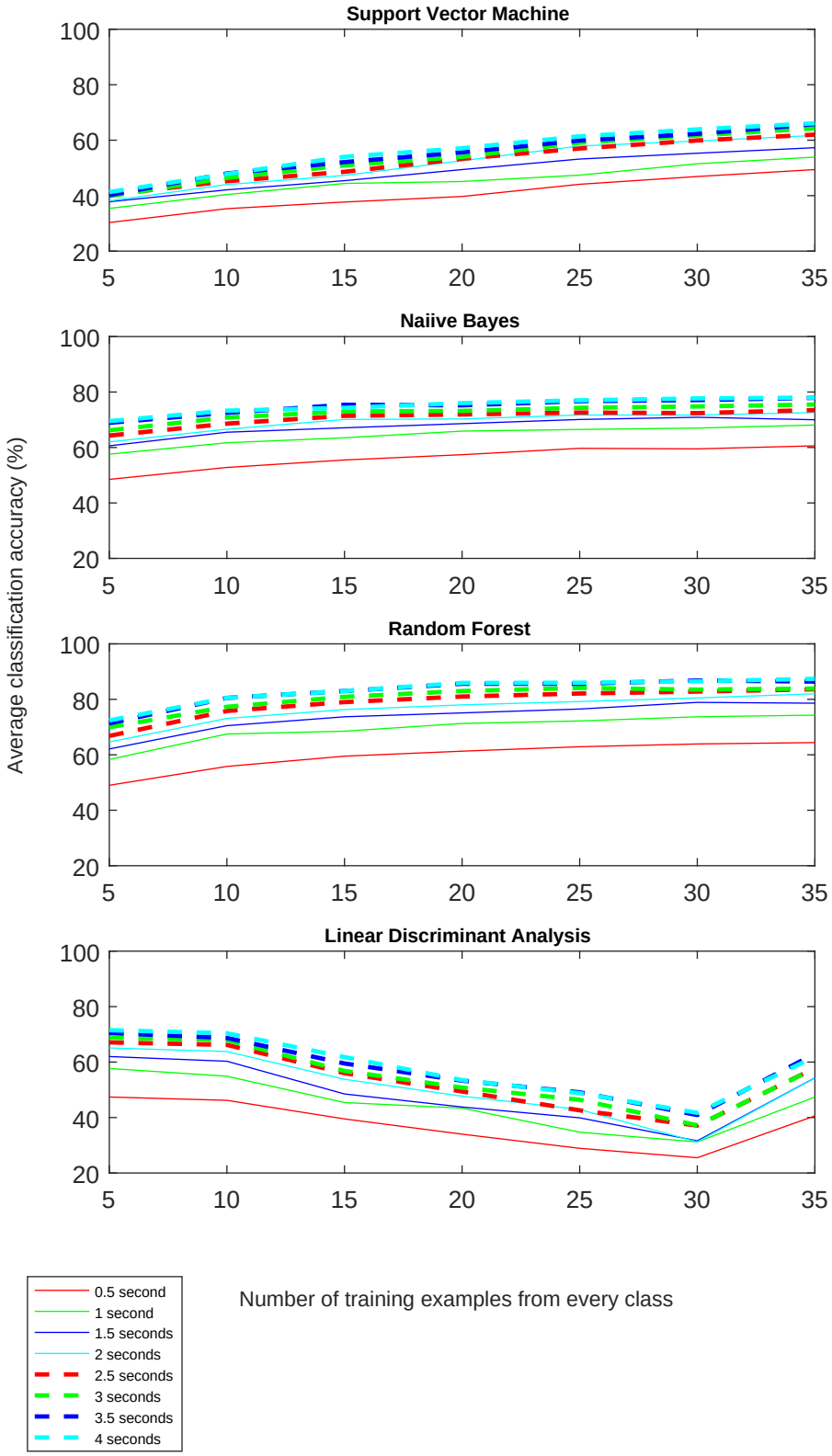

Fig. 5.3 Average classification accuracy (%) of the fixed time separated trials data in classifying 5 imagined words, using different classifiers, when different training sizes and different time frames are used

5.3.3 The relation between repetitions order and classification accuracy

In the mouse click separated trials data, 60 trials were recorded in two blocks: 40 and 20 trials for every word. 10-fold cross validation was applied where the portion of training and testing data from each block is proportional to the size of the block. From Table 5.5, the maximum average accuracy achieved is 62.94% using RF and total number of training 270 trials. In comparison to Table 5.1, using data from the same block and 175 trials, 71.65% average classification accuracy using RF can be obtained. Moreover, in comparison to Figure 5.2, 62.50% using RF is achieved using 50 total training trials. However, having each word recorded in one separate block leads to a high temporal correlation in EEG patterns across different words. Thus, recording using sub-blocks or random representation is more representative as the temporal correlation is reduced in EEG patterns of each class. This issue has been investigated in (Porbadnigk et al., 2009).

Table 5.5 10-folds average classification accuracy to classify between five words for mouse click separated data; using training and testing data mixed from two different blocks for each word.

Subject	SVM	NB	RF	LDA
S1	52.33	50.67	68.00	48.00
S2	51.60	41.00	53.00	50.67
S3	41.67	46.67	57.33	46.33
S4	53.67	57.33	72.67	50.67
S5	63.33	66.33	82.67	59.00
S6	29.33	31.33	46.33	38.33
S7	58.33	44.33	73.67	56.67
S8	49.67	49.00	52.33	46.67
S9	41.00	49.33	59.33	38.00
S10	40.67	37.33	64.33	43.00
Average	48.12	47.30	62.94	47.70

Feature	SVM	NB	RF	LDA
DWT	56.04	64.57	71.65	52.16
Imagination length	37.25	36.95	30.55	36.95
DWT and Imagination length	59.98	67.26	73.90	50.01

Table 5.6 10-fold average classification accuracy (%) using different features for mouse click trials separation data by using 35 training trials for every word.

5.3.4 The effect of imagination time

In the mouse click trials separation data, the participant determined the start and the end of the imagination trial using mouse clicks. Figure 5.4 shows the average time needed for each participant to imagine every word. Across subjects, the average imagination length for the five words are: 1.8, 1.5, 1.3, 1.5, and 1.6 seconds for the words: *"left"*, *"right"*, *"up"*, *"down"*, and *"select"* respectively. As shown in Tables 5.6 and 5.7, adding the imagination length as an extra feature improves the average classification accuracy for SVM, NB, and RF classifiers by an average of 2.25% - 3.94%. That means the imagination length is possibly an effective feature for classifying the words. However, applying t-test shows that for none of the classifiers this improvement is statically significant. Importantly, the examination of how the imagination length for each word may vary across blocks recorded needs to be investigated because the learning curve might affect how the subjects perform the imagination task.

In Table 5.8 the effect of having subject specific time frame has been examined. This time frame was adapted by reducing the fixed time frame to a length that is approximately equal to the maximum average length the participant needed in mouse click separated imagination for any of the imagined words (from Figure 5.4). In comparison to the classification accuracies in table 5.1, for fixed time frame separated data, the differences are statistically significant only for all the four classifiers. This also approves what has been explained in Section 5.3.2 that long fixed time frame provides low frequencies in the extracted time window to help in distinguishing EEG patterns related to speech. In future work more investigations can be performed by making the length adaptation for each word separately.

Table 5.7 10-folds average classification accuracy (%) to classify between five words for mouse click separated data; using DWT and word length as classification feature; bold means the maximum accuracy for this subject.

Subject	SVM	NB	RF	LDA
S1	67.22	74.04	**87.28**	54.53
S2	53.45	60.85	**65.67**	45.50
S3	52.87	67.19	**71.43**	54.56
S4	60.82	**79.80**	78.22	55.53
S5	90.47	93.65	**95.79**	68.80
S6	41.35	46.43	**56.58**	31.78
S7	66.67	53.48	**70.94**	47.57
S8	58.63	57.60	**65.53**	45.99
S9	48.07	67.69	69.82	38.77
S10	60.26	71.9	**77.75**	57.11
Average	59.97	67.26	73.90	50.01

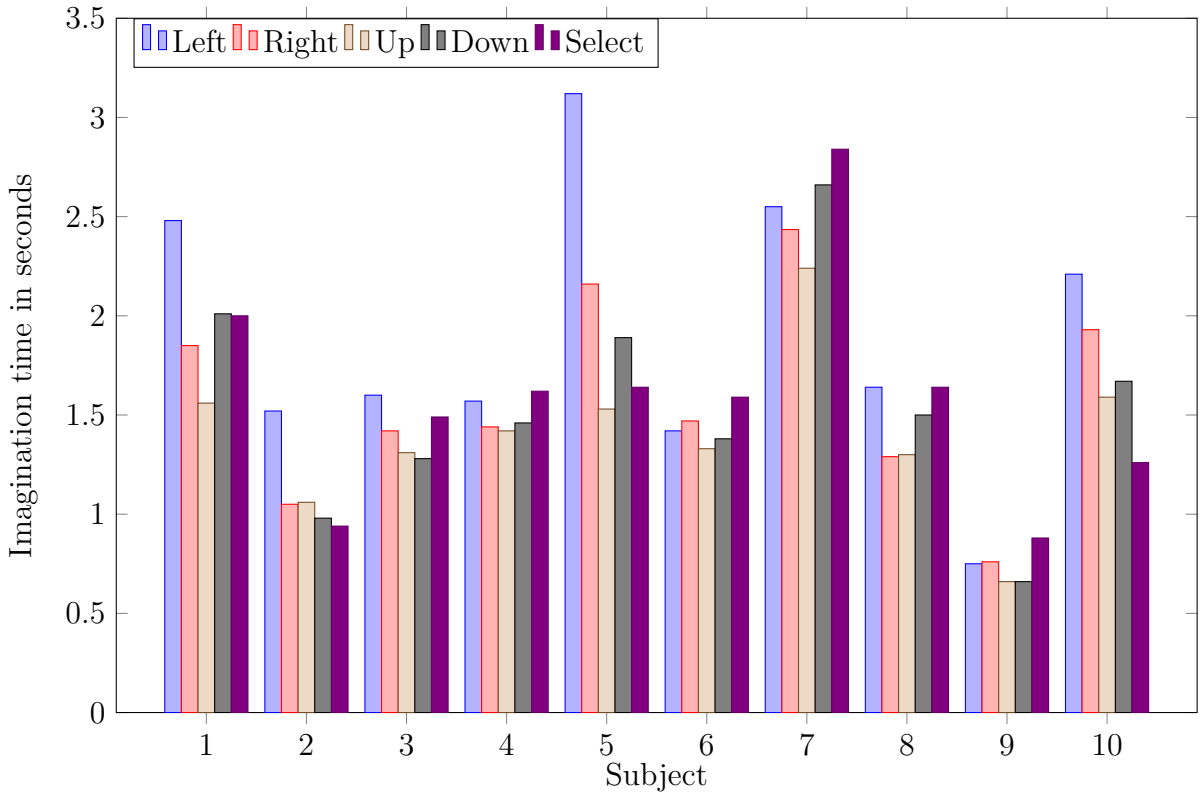

Fig. 5.4 Average imagination time in second using 40 trials from every word

Table 5.8 10-folds average classification accuracy to classify between five words where for each subject the time frame is adopted to the average time frame for the word with the maximum length in mouse click separated trials; arrows show the increase/decrease in classification compared to results in Table 5.1; t-test between the results in this table and Table 5.1

	Average length of the	Fixed time frame			
Subject	longest word (in seconds)	SVM	NB	RF	LDA
S1	2.5	54.79 (6.5 ↓)	67.87 (9.47 ↓)	85.97 (0.45 ↓)	45.32 (4.39 ↓)
S2	1.5	60.74 (8.1 ↓)	64.24 (18.65 ↓)	72.84 (11.58 ↓)	53.89 (13.45 ↓)
S3	1.5	54.79 (6 ↓)	66.79 (5.58 ↓)	69.74 (19.21 ↓)	56.76 (1.53 ↓)
S4	1.5	55.74 (12.52 ↓)	85.32 (5.65 ↓)	91.42 (7.05 ↓)	62.24 (12.07 ↓)
S5	3	85.29 (2.21 ↓)	86.21 (0.63 ↑)	88.08 (4.38 ↓)	47.91 (7.58 ↓)
S6	1.5	43.34 (12.03 ↓)	64.34 (12.55 ↓)	68.29 (12.1 ↓)	46.18 (5.66 ↓)
S7	3	87.45 (0.03 ↑)	79.95 (3 ↓)	91.92 (2.03 ↓)	75.34 (1.13 ↓)
S8	1.5	59.82 (8.55 ↓)	63.34 (3.44 ↓)	79.95 (3.52 ↓)	54.32 (0.55 ↓)
S9	1	71.82 (12.12 ↓)	91.47 (5.53 ↓)	92.42 (4.58 ↓)	65.37 (18.58 ↓)
S10	2.5	38.16 (2.02 ↓)	55.24 (4.52 ↓)	65.26 (8.53 ↓)	32.63 (7.58 ↓)
Average	1.8	58.51 (7.02 ↓ 0.003 ↑)	70.95 (6.84 ↓ 0.063 ↑)	79.76 (7.34 ↓ 0.00 ↑)	54.67 (7.25 ↓ 0.76 ↑)
T-test		✓(0.0008)	✓(0.0046)	✓(0.0027)	✓(0.04)

5.4 Conclusions

This chapter addresses several questions related to the design of unspoken speech studies in a block recording mode where the trials separated using mouse click and fixed time frame. First, the relation between training size (5-35 trials) and the classifier performance using the dataset collected by imagining five different words and four classifiers was examined. Due to the limitation in the collected number of trials for each word, observing any saturation in the classification across different number of training trials was difficult. However, the results show that the rate of improvement in accuracy gets very small when moving from 25-35 training trials for each class. In contrast, this improvement increased sharply when the training size was increased from 5-15 trials for every class. For all training sizes and both data separation methods, random forest (RF) classifier provides the highest average classification accuracy. However, after mixing data recorded in two different time the classification accuracy resulting from random forest significantly dropped.

Second, for fixed time separation, it has been found that the longest time frame provides DWT features that lead to best results. 3.5-4 seconds gives the maximum average accuracy. Third, the system was trained using data from two blocks recorded in the same session but more training trials needed to get equivalent performance to

classification using one block. Finally, the use of mouse click to separate the trials showed that the imagined speech rate was less than real spoken speech. The participants needed 1.8 seconds on average to imagine the longest word even after removing the time needed to do mouse click (on average 100 ms for male adults (Komandur et al., 2008)).

5.5 Summary

Studies on recognising unspoken speech with the use of EEG signals vary in their designs. The participants are either asked to imagine unspoken speech within a specific time frame, or alternatively indicate the start and end of the imagined speech. Optimizing the length and training size of imagined speech is important to improve the rate and speed of recognizing unspoken speech in on-line applications. This chapter examined the recognition of unspoken speech in block recording mode as well as studying the experimental parameters to improve the classification accuracy.

In this study, EEG data was recorded when the participants performed unspoken speech of five words using two technologies: (1) marking the start and end of the trial by using mouse clicks and (2) performing the imagination in a four-second fixed time window. Four classifiers were trained in all experiment parts: support vector machine (SVM), naive bayes (NB), random forest (RF), and linear discriminate analysis (LDA). The results show that the best time frame is 3.5-4 seconds length. Moreover, the increase in training size improve the average classification accuracy. However, this improvement becomes slight between 125-175 total training trials. The training data can be recorded in parts, however, the required training size should be increased to have better classification accuracy. In all experiment parts, random forest classifier shows better results among the other classifiers.

Chapter 6

Dynamic Time Warping in the Recognition of Imagined Speech

A time series is a sequence of data points arranged in chronological order and EEG data is an example of multi-dimensional time series. Time domain (TD) analysis stems from the desire to understand the signal in its original state without having to represent it in terms of frequency. TD analysis is the most direct way to examine EEG signals.

The limited TD analysis is clearly visible in the literature review presented in Chapter 3, which focuses on speech imagination. Although few studies have been conducted on the recognition of imagined speech using time-domain features, they mostly failed to consider temporal variations of imagined speech. Such variations are caused by a number of factors including differences in the start time, the speed of imagining the pronounced words, and variations between the imagined words and can significantly degrade the classification results. Thus, the main motivation of this chapter was to propose and evaluate a novel framework based on DTW to classify EEG signals by minimising the temporal variations across EEG trials from the same class.

With the DTW technique, two discrete time series are compared by warping them against each other. In doing so, any temporal differences between the two sequences will be minimised. The warping process determines when and where the examined time series should be expanded or compressed in time to find the most

representative difference between them. Unlike Euclidean distance and similarity-measurement technique of cross-correlation, the DTW technique does not require normalisation to make the two compared time series equal in length. DTW is commonly used in the field of speech processing (Sakoe and Chiba, 1978). Thus, DTW is a powerful algorithm for analysing a variety of time series, such as audio, video or images (Rakthanmanon et al., 2012).

This chapter describes a novel framework for feature extraction that was developed based on the DTW algorithm. The proposed framework represents the first use of DTW in the context of imagined speech recognition. The evaluation was performed using the recorded EEG data that were presented in Chapter 5. The proposed DTW framework was used to compare three types of features: TD features (maximum cross-correlation (MaxCC), statistics of EEG signals), modified DTW approaches from the literature and time-frequency features. The time-frequency features included energy calculated from discrete wavelet transform (RWE-DWT), statistics of discrete wavelet transform (statistics-DWT). Moreover, common spatial patterns (CSPs) was examined as spatial s. The evaluation involved discriminating between imagined speech versus silence, and discriminating between five imagined words.

The remainder of this chapter is structured as follows. An overview of DTW and how the technique works is presented in Section 6.1. Examples from the literature on how DTW was used in BCI studies are provided in Section 6.2. The proposed framework is presented in Section 6.3. The algorithms that were compared to the proposed framework and experiments that were conducted using the proposed features are discussed in Section 6.4. Section 6.5 presents the used data, feature extraction, and the classification algorithms. The results of the experiments are discussed in Section 6.6. Finally, several conclusions are presented in Section 6.7. Section 6.8 summarises this chapter.

6.1 Dynamic time warping

The main aim of DTW is to compare two time series by reducing the time distortion and finding the best alignment between them. Assume we have two time series of different lengths, namely $\boldsymbol{A}$ of length n and $\boldsymbol{B}$ of length m, where

$$\boldsymbol{A} = [a_1, a_2, \cdots, a_i, \cdots, a_n], \tag{6.1}$$

and

$$\boldsymbol{B} = [b_1, b_2, \cdots, b_j, \cdots, b_m]. \tag{6.2}$$

First, the distance between each point a_i in $\boldsymbol{A}$ and b_j in $\boldsymbol{B}$, $\mathrm{d}(a_i, b_j)$, is calculated using a suitable measure, such as the Euclidean distance. As a result, the distance matrix $\boldsymbol{D}^{n\times m}$ (also known as a cost matrix) is obtained in which each element $\boldsymbol{D}(i,j)$ represents the distance between a_i and b_j.

The next step is to map the elements of $\boldsymbol{A}$ and $\boldsymbol{B}$ through the matrix $\boldsymbol{D}$ by finding an optimum warping path such that the cumulative distance between the two time series is minimized. A warping path $\boldsymbol{P}$ belongs to a set of warping paths Ω in matrix $\boldsymbol{D}$, and is denoted as:

$$\boldsymbol{P} = [p_1, p_2, \cdots, p_y, \cdots, p_Y] \;\; (length\ Y), \tag{6.3}$$

where any element of $\boldsymbol{P}$ is defined as $p_y = \boldsymbol{D}(i,j)_y$ and $\mathbf{max}\ (m,n) \leq Y < m+n-1$. The total DTW distance between the two time series is given by the optimum $\boldsymbol{P}$ such that:

$$\mathbf{DTW}(\boldsymbol{A}, \boldsymbol{B}) = \min_{P\in\ \Omega} \left[\frac{1}{Y} \sqrt{\sum_{y=1}^{Y} p_y^2} \right], \tag{6.4}$$

where Ω is the set of all paths. The object function in (6.4) needs to satisfy constraints: continuity, monotonicity, and boundary Sakoe and Chiba (1978):

- Continuity: The path advances one step at time to one of the adjacent cells. Given $p_y = \boldsymbol{D}(x,z)$, $p_{y-1} = \boldsymbol{D}(x',z')$ where $x - x' \leqslant 1$ and $z - z' \leqslant 1$.

- Monotonicity: This condition guarantees that the path will not turn back on itself. Given $p_y = \boldsymbol{D}(x,z)$, $p_{y-1} = \boldsymbol{D}(x', z')$ where $x - x' \geqslant 0$ and $z - z' \geqslant 0$.
- Boundary constraint: start and end points of the warping path have to be the very first and last points of the given time signals, where $p_1 = \boldsymbol{D}(1,1)$ and $p_Y = \boldsymbol{D}(n, m)$.
- Slope constraint: The slope of the warping path should be nor too gentle or too restricted to avoid unrealistic alignment between too short or too long patterns. The slope can be between 0 and to ∞ (Sakoe and Chiba, 1978).

Dynamic programming can be used to reduce the computational cost associated with finding the optimum warping path (Bellman, 2013). The cumulative distance between two points $a_i \in \boldsymbol{A}$ and $b_j \in \boldsymbol{B}$, $\boldsymbol{cd}(a_i, b_j)$, can be calculated using the following recursion:

$$\begin{aligned} \boldsymbol{cd}(a_i, b_j) = \mathbf{min}\ [&\boldsymbol{cd}(a_{i-1}, b_j), \boldsymbol{cd}(a_{i-1}, b_{j-1}), \\ &\boldsymbol{cd}(a_i, b_{j-1})] + \mathbf{d}(a_i, b_j), \end{aligned} \tag{6.5}$$

where $\boldsymbol{cd}(a_1, b_1)$=$\mathbf{d}(a_1, b_1)$. Hence, the DTW distance between two time series is equal to the cumulative distance at the end of the optimal path:

$$\mathbf{DTW}(\boldsymbol{A}, \boldsymbol{B}) = \boldsymbol{cd}(a_n, b_m) \tag{6.6}$$

There are two main types of DTW based on how the two times series are aligned: symmetric and asymmetric DTW. In symmetric DTW the two series are transformed on to one temporal time-axis. In asymmetric DTW the time normalization is performed to have one temporal pattern aligned to that of the other (Sakoe and Chiba, 1978). In the current study, symmetric DTW was used.

6.2 DTW using EEG and ECoG

Although DTW was originally developed primarily for speech recognition systems, several studies have utilised DTW for biosignal analysis. In the brain-signal domain, previous studies have examined EEG and ECoG in several research contexts. The following are examples of these studies, presented in chronological order.

Chaovalitwongse and Pardalos (2008) integrated DTW with SVM as a kernel function to distinguish between brain signals in normal people and pre-seizure patients. The results showed that SVM with DTW was superior to the regular SVM in the classification. A pair-wise kernel function was calculated for each electrode. In this case, DTW was used on the inner products of the kernel function to find the optimal path.

For spoken sentence classification from ECoG, Zhang et al. (2012) used DTW for utterance matching. First, the authors calculated templates by realigning high-gamma features from ECoG with the corresponding output sound using DTW. These alignments were then averaged to create a template for each of the two sentences. When any new input arrived, the correlation between the input and the templates was calculated and classified using Fisher discriminant analysis. The results showed better classification than SVM.

Karamzadeh et al. (2013), combined DTW with quality threshold (QT) clustering to trace brain areas that are functionally connected during specific tasks. The tasks included both visual and audio tasks. EEG data were segmented into temporal windows based on well-known intervals and activities. DTW was then computed for each EEG segment. The channels were then classified based on the similarity of their behaviour.

Zoumpoulaki et al. (2015) utilised DTW as an ERPs latency measurement. The researchers argued that DTW provides the advantage of measuring the latency by comparing two time series to overcome the problems of older methods, which depended on one ERPs pattern only. The authors also used DTW to measure the latency of all the available points instead of comparing one point only, which helped in determining when and where the latency occurred. The experiments were performed on artificial

and real EEG data and focussed on the channels Cz (to analyse the P1 component) and Pz (to analyse the P3 component).

Gui et al. (2015) used DTW to compare EEG patterns in order to achieve user identification. The focus of their study was on the representation of four channels in the left superior temporal lobe (Pz, O1, O2, and Oz). The participants were asked to read a list of unconnected text. The list contained four groupings of words: 75 words, 75 pseudo-words, 75 illegal strings, and 150 instances of their own names. The recording was performed twice: once for training and once for testing.

The main goal of Martin et al. (2016) study was to classify ECoG for six words in three different modes: imagination, listening, and voicedly speaking. The researchers selected these words to reflect variations in the number of syllables, acoustic features, and semantics. The authors used DTW in an SVM classifier kernel function as a form of non-linear alignment between trials. The researchers found listening and speaking classification outperformed imagination classification, as imagination was not easy to align because of the difficulty in determining the start and end of tasks.

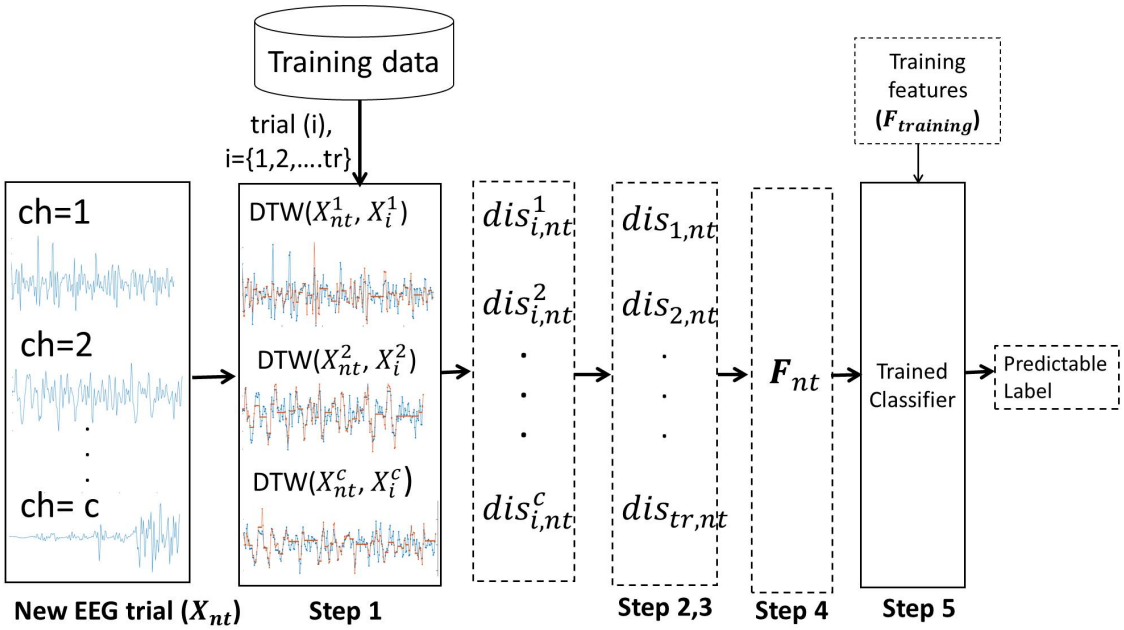

Fig. 6.1 Architecture of the DTW-based framework for EEG data classification

6.3 Proposed DTW-based framework for feature extraction

6.3.1 Computing training features using the proposed DTW framework

Let's assume the i^{th} EEG training trial is presented as $\boldsymbol{X}_i \in \boldsymbol{R}^{n \times c}$ where n is the number of samples, c is the number of channels. $i \in \{1, 2, 3, \cdots, tr\}$ denotes the trial number where tr is the training data size. The features presenting $\boldsymbol{X}_i$ can be calculated as follows:

- *Step 1*: Compute DTW-based distance, dis, between the trial $\boldsymbol{X}_i^{ch} \in \mathbf{R}^{n \times c}$ and each $\boldsymbol{X}_j^{ch} \in \boldsymbol{R}^{m \times c}$. The distance is computed between the samples from each channel ch, $\{ch = 1, 2, 3, \cdots, c\}$, separately as

$$dis_{i,j}^{ch} = \mathbf{DTW}(X_i^{ch}, X_j^{ch}). \tag{6.7}$$

- *Step 2:* Computing the total distance between $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ by averaging the distances across all the channels,

$$dis_{i,j} = \frac{1}{c} \sum_{ch=1}^{c} dis_{i,j}^{ch}. \tag{6.8}$$

- Step 3: The feature vector for X_i can be presented as

$$\boldsymbol{F}_i = [dis_{i,1}, dis_{i,2}, \cdots, dis_{i,tr-1}, dis_{i,tr}], \tag{6.9}$$

  where $dis_{i,i} = 0$.

- Step 4: The final feature matrix presenting all training features is:

$$\boldsymbol{F}_{training} = [\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3, \cdots, \boldsymbol{F}_i, \cdots, \boldsymbol{F}_{tr}]. \tag{6.10}$$

- Step 5: The classification algorithms are trained using the feature matrix $\boldsymbol{F}_{training}$ and the corresponding class labels.

6.3.2 Classifying a new EEG trial using the proposed framework

Figure 6.1 illustrates the steps for classifying a new EEG trial. For a new test trial, $\boldsymbol{X}_{nt} \in \boldsymbol{R}^{s \times c}$, the steps for computing testing features are similar to the steps for computing training features. In step 1, the DTW distance is computed between $\boldsymbol{X}_{nt}^{ch}$ and each training trial from the corresponding channel. In step 2, the total distance between $\boldsymbol{X}_{nt}$ and each training trial is computed by averaging the distances across all the channels. In step 3, the feature vector of the new trial, $\boldsymbol{F}_{nt}$, is created by concatenating the total distances between $\boldsymbol{X}_{nt}$ and the training trials. In step 4, testing features, $\boldsymbol{F}_{nt}$, is fed into the trained classifier in order to classify the testing trial.

6.4 Methodology

The main aim of the experimental methodology used in this study is to prove the usability of the proposed DTW feature extraction framework. The evaluation aimed to: (1) compare the framework with TD features, (2) examine several modifications to it and (3) compare it with time-frequency features. Figure 6.2 lists the evaluation methods used in the study. As the data collection evolved, two different methods were used for trial separation: mouse clicks and fixed time frame separations.

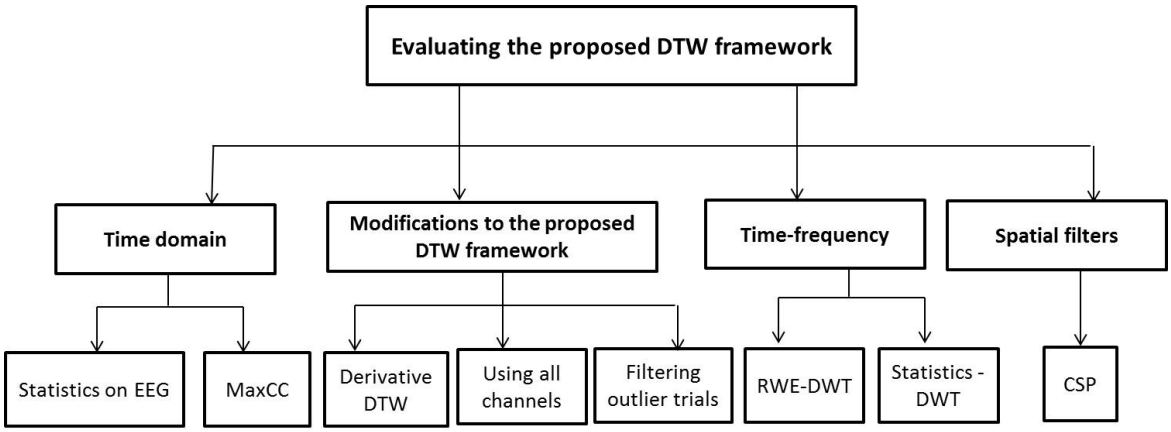

Fig. 6.2 The proposed methods to be compared with the proposed DTW framework

6.4.1 Comparison with time-domain feature extraction algorithms

Statistics of EEG data

Statistics of EEG data was proposed one of time-domain features to be compared with DTW-based features. For each EEG trial $\boldsymbol{X}_i$ from training or testing data, statistics of the samples in each EEG channel separately were calculated. Four statistical measures were used: mean, SD, root mean square (RMS), and sum of EEG values (SUM) (see Chapter 4 section 4.2.6).

Maximum cross-correlation (MaxCC)

Cross-correlation (**CC**) is an algorithm to examine the time delay that exists between two time series. The output of the **CC** is called a cross-correlogram (***CCo***).

In the context of EEG data research, several studies have utilised **CC** as a similarity measure or a feature-extraction method. For example, in (Krishna et al., 2016), statistics of ***CCo*** coefficients were computed from EEG signals and used as features to classify different motor imaginations. In (Kumagai et al., 2017), **CC** was proposed as an approach to determine the degree to which a listener is familiar with the music they are exposed to. This was performed by computing **CC** between the EEG signal and the music envelope. In (Bhavsar et al., 2018), **CC** was used to investigate the relationship between the variability in EEG signals and electrode positioning. For classifying imagined words using EEG, maximum linear cross correlation was used as one of the feature sets in (Qureshi et al., 2018).

Let's assume $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ are the signals, m is the signals' length and ϕ is the time shift parameter $\phi = \{-m+1, ... -3, -2, -1, 0, 1, 2, 3.....m+1\}$. If $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ do not have the same length, the shortest signal is extended with zeros.

The following equation explains the computation of ***CCo***:

$$\mathbf{CC}(X_i, X_j, \phi) = \begin{cases} \sum_{t=0}^{m-\phi-1} \boldsymbol{X}_{i_{t+\phi}} \cdot \boldsymbol{X}_{j_t}, & if\ \phi \geqslant 0 \\ \mathbf{CC}(\boldsymbol{X}_i, \boldsymbol{X}_j, -\phi) & otherwise. \end{cases} \tag{6.11}$$

A variety of statistical features have been extracted from the ***CCo*** in the literature, including mean, median, SD, maximum, mode, and minimum. In the present study, after comparing several statistical measures on ***CCo***, the maximum absolute value of the **CCo** was selected as the classification feature.

$$\mathbf{MaxCC}(\boldsymbol{CCo}) = \mathbf{max}(|\boldsymbol{CCo}|), \tag{6.12}$$

The feature-extraction procedure is similar to the steps followed for the DTW feature-extraction steps described in section 6.3. Instead of computing DTW distance in step 1, $MaxCCo$ was computed using equation.6.12 for each channel separately as explained in the following equations:

$$MaxCCo_{i,j}^{ch} = \mathbf{MaxCC}(X_i^{ch}, X_j^{ch}), \tag{6.13}$$

$$MaxCCo_{i,j} = \frac{1}{c}\sum_{ch=1}^{c} MaxCCo_{i,j}^{ch}; \tag{6.14}$$

6.4.2 Comparison with time-frequency algorithms

Features based-on DWT coefficients

Discrete wavelet transform (DWT) was explained in details in section 5.2.2. In this chapter in addition to statistics of DWT coefficients (statistics-DWT), relative wavelet energy (RWE) was also calculated from the coefficients as another feature set. RWE was calculated for the coefficients in the detailed levels and the approximation level by comparing the energy contribution of each decomposition level to the total wavelet energy (Guo et al., 2009). This procedure was previously applied in (González-Castañeda et al., 2017) for classification of imagined speech.

As the number of decomposition levels was five, this resulted in having five detailed levels, d_l where $l \in \{1, 2, \cdots, 5\}$, and one approximation level a. Moreover, each level has a set of coefficients CO, $co \in \{1, 2, \cdots, CO\}$. In this study the number of coefficients in each level for each trial differs due to the difference in the trials length in the case of mouse click separated data. Next, the energy in level l, E_l, for detailed levels and the approximation level can be computed as:

$$\boldsymbol{E}_l = \begin{cases} \sum_{CO} |d_{l,co}|^2, & if\ l \leq 5 \\ \sum_{CO} |a_{co}|^2, & otherwise. \end{cases} \tag{6.15}$$

Using the following equation, the total RWE of level l can be calculated as:

$$\mathbf{RWE}_l = \frac{\boldsymbol{E}_l}{\boldsymbol{E}_{total}} \quad \text{where} \quad \boldsymbol{E}_{total} = \sum_{l=1}^{L} \boldsymbol{E}_l. \tag{6.16}$$

6.4.3 Comparison with Common Spatial Pattern (CSP)

Common spatial patterns have been described in Section 4.2.6. CSPs was designed in the literature to capture the spatio-temporal information from the poor spatial resolution EEGs. However, in this chapter classical CSP (not filter-bank) was used as the results are compared with the proposed DTW framework in a single frequency range.

6.4.4 Modifications to the proposed framework

Derivative dynamic time warping (DDTW)

Keogh and Pazzani (2001) proposed derivative dynamic time warping (DDTW) as a solution to problems that might arise in dynamic time warping alignment. Among these problems are what the authors call 'singularities', which occur when the warping path represents strong variability in the Y-axis of the time series by warping of the X-axis. As a result, a single point becomes warped to a large subsection of the other time series. In addition to encountering singularity, the authors found that DTW sometimes failed to perform correct warping because of slight variations between features (for example the peak, valley, inflection point, and plateau).

DDTW and DTW have two main differences. First, instead of using the raw EEG data of the two time series, the derivatives of the time series are used. Second, instead of using Euclidean distance, the square of the Euclidean distance between the derivatives of the two time series is calculated.

For an EEG trial $\boldsymbol{X}_i \in \boldsymbol{R}^{n \times c}$ where n is the number of samples in $\boldsymbol{X}_i$, $\{1, 2, 3, \cdots, n\}$, c is the number of channels, $ch = \{1, 2, 3, \cdots, c\}$, the derivative of $\boldsymbol{X}_i^{ch}$ (according to

(Keogh and Pazzani, 2001)) is:

$$\frac{(\boldsymbol{X}_{i_s}^{ch} - \boldsymbol{X}_{i_{s-1}}^{ch}) + ((\boldsymbol{X}_{i_{s+1}}^{ch} - \boldsymbol{X}_{i_{s-1}}^{ch})/2)}{2} \tag{6.17}$$

Filtering outlier trials to improve the proposed framework

In the proposed DTW framework, the distance is computed between each training trial and the rest of training trials. A trial is considered an outlier when the distance between the trial and the rest of the trials from the same class is higher than the distance between the trial and the trials from the other classes. This check is performed for the training trials to ensure consistency in distances. The following steps were applied to compute the filtering of the outlier trials:

1. The average DTW distance between each training trial $\boldsymbol{X}_i \in Class_A$, $A = \{1, 2, 3, 4, 5\}$, and each $\boldsymbol{X}_j \in Class_B$, $B = \{1, 2, 3, 4, 5\}$, $Class_A \neq Class_B$, is computed. This followed the same steps for computing training features as in section 6.3.

2. For each $\boldsymbol{X}_i$, the average of the average DTW distances is computed. For each $\boldsymbol{X}_i$ one value represents average from all out of class training trials. As a result, each class has a list of distances equal to the number of training trials related to that class.

3. For each list resulted from the previous step, the median value is computed.

4. The difference from the median is computed for each value in each list.

5. The different values are listed in descending order as a representation of the trials' consistency; trials with longer average distance from the training trials from the other classes are more representative of that class.

Dynamic time warping for words classification using all channels

In the proposed DTW framework described in Section 6.3, the distance between two EEG trials $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ is computed from each channel independently. In computing Euclidean distance in one-dimension space, the distance between each sample in $\boldsymbol{X}_i^{ch}$ and in X_j^{ch}:

$$\left|\boldsymbol{X}_{i_n}^{ch} - X_{j_m}^{ch}\right|, \tag{6.18}$$

where $n = \{1, \cdots, \mathbf{length}(\boldsymbol{X}_i)\}$, and $m = \{1, \cdots, \mathbf{length}(\boldsymbol{X}_j)\}$.

In the classical DTW approach used for audio-speech recognition, the distance between two utterances is computed using all frequency channels together in such a way that the Euclidean distance is computed in 2-D. This approach has been examined in comparison to the proposed framework. The Euclidean distance between $\boldsymbol{X}_{i_1}$ and $\boldsymbol{X}_{j_1}$:

$$\sqrt{(\boldsymbol{X}_{i_1}^1 - \boldsymbol{X}_{j_1}^1)^2 + (\boldsymbol{X}_{i_1}^2 - \boldsymbol{X}_{j_1}^2)^2 \cdots + \boldsymbol{X}_{i_1}^{ch} - \boldsymbol{X}_{j_1}^{ch})^2}, \tag{6.19}$$

where $ch = \{1, 2, \cdots, 12\}$.

6.5 Experiment

6.5.1 Data collection

EEG data set was described in Chapter 5 section 5.1. This data involved the imagination of five different words: *"left"*, *"right"*, *"up"*, *"down"* and *"select"*. The data were recorded with two different data separation methods: fixed time frame and mouse clicks to separate between trials. For mouse click separated data, 60 trials were recorded for each subject during two blocks. For fixed time separated data, 40 trials were recorded for each subject in one block.

6.5.2 Feature extraction

For DTW-based features (as explained in Section 6.3), the DTW distance between each pair of training trials was used as a training features. For testing the distance between each testing trial and each training trial was used as the test features. Thus, the number of extracted features per trial was equal to the number of the train trials. For MaxCC features, because the same feature extraction strategy as DTW-based features was applied, the number of the total number of features were the same as the one for the DTW-based features.

For the RWE-DWT features, in this study the number of decomposition levels was five, RWE was then calculated for all the five detailed levels and the approximation level. Because 12 EEG channels were involved in this study, with five detailed and one approximation levels, a total of 72 features were extracted per trial (i.e. 12 channels $\times$ 6 decomposition levels). For DWT-statistics as explained in 5.2.2, the number of features was 144.

For statistics of EEG data, a total of 48 features were extracted from each trial (4 statistical measures $\times$ 12 channels). In CSP, the number of features is determined by the number of rows selected from the designed spatial filters. Usually, a specific number of rows from top and the bottom of the spatial filter matrix is selected. In this experiment, the number of selected spatial signals was four.

6.5.3 Classification

Four different classifiers were trained to classify the five words: support vector machine (SVM), naïve bayes (NB), random forest (RF), and linear discriminant analysis (LDA). In this experiment for RF, a total of 50 trees were used in our experiment, and the number of nodes in each tree was calculated as $\log_2(Numberoffeatures + 1)$ as in González-Castañeda et al. (2017). Details about these classifiers were explained in Chapter 4 section 4.2.7.

In this study, the classification model was subject-specific, and 10-fold cross-validation was used for evaluation. For binary classification between silence and unspoken speech, only the data from the first block of all the five words were used. The total number of unspoken speech trials (i.e. imagined words) was 200, and the total number of silence trials was 195 (the first silence trial for each word was not counted). For classifying between the words, the two recorded blocks for each word were used (i.e. 40 trials and 20 trials respectively), the data were mixed in equal percentages from each block for each word in the training and testing sets. In each fold, 270 training trials were used (54 trials from each class), and 30 trials were used for testing (6 trials from each class). Finally, paired t-tests were used to compare different feature sets in terms of classification accuracy.

6.6 Analysis and Results

The general aim of the following experiments is to verify the proposed framework in comparison with time-domain features, some modifications to the framework, time-frequency features, and spatial filters. Also, there are two data sets that are varying based on the word separation technique (mouse click and fixed time frame). Table 6.1 lists the experiments that were designed to cover all the evaluation aims.

Table 6.1 Experimental design to evaluate the proposed DTW framework

Experiment/ section number	Aim	Compared technique	Data
1 (section 6.6.1)	Classifying speech versus non-speech	MaxCC, and statistics of EEG	mouse click separated data
2 (section 6.6.2)	Classifying imagined words	MaxCC, and statistics of EEG	Mouse click separated data
3 (section 6.6.3)	Classifying speech versus non-speech	MaxCC, statistics of EEG	Fixed time separated data
4 (section 6.6.4)	Classifying imagined words	MaxCC, and statistics of EEG	Fixed time separated data
5 (section 6.6.5)	Classifying imagined words	Derivative DTW	Mouse click separated data
6 (section 6.6.6)	Classifying imagined words	Using all channels	Mouse click separated data
7 (section 6.6.7)	Classifying imagined words	Filtering outlier trials	Mouse click separated data and fixed time separated data
8 (section 6.6.8)	Classifying speech versus non-speech	DWT-statistics, and RWE-DWT	Mouse click separated data and fixed time separated data
9 (section 6.6.9)	Classifying imagined words	DWT-statistics,and RWE-DWT	Mouse click separated data and fixed time separated data
10 (section 6.6.10)	Classifying speech versus non-speech	CSP	fixed time separated data

6.6.1 Classifying imagined speech versus silence for mouse click separated data using time-domain features

The first part of the evaluation was conducted to examine the effectiveness of the proposed framework in classifying unspoken speech trials (from all five words). It also sought to assess the rest trials (i.e. the time between clicks before imagining a word). Table 6.2 compares the 10-fold classification results for unspoken speech and silence using the four different sets of TD features and the four classifiers. Using the DTW features, the RF classifier provided the highest average classification accuracy (72.35%). However, this average result was only 0.32% higher than that of the LDA classifier. The second-best set of features was MaxCC, which had an average classification accuracy of 68.96% with the RF classifier. This result was 3.39% lower than the classification accuracy of RF using DTW. The pairwise t-test classification results of the DTW and MaxCC features were not statistically significant ($p = 0.14$) with the RF classifier (Table 6.3). For LDA, DTW was found to outperform MaxCC, with a 6.18% average improvement. This performance tended to be statistically significant ($p = 0.059$). Moreover, there are no important findings to highlight regarding statistics of EEG.

Table 6.2 Average 10-fold cross-validation results (%) of classifying unspoken speech versus silence for mouse click separated data using three time-domain feature-extraction methods across four classifiers; the best result for every subject is in bold

Subject	DTW				MaxCC				Statistics of EEG			
	SVM	NB	RF	LDA	SVM	NB	RF	LDA	SVM	NB	RF	LDA
1	31.94	76.66	81.24	81.49	49.49	76.15	**82.99**	81.22	50.51	50.77	62.39	52.04
2	88.48	87.05	93.64	**95.91**	49.49	78.12	85.25	76.40	50.51	50.78	77.15	50.51
3	50.25	58.08	**58.66**	51.75	48.72	55.04	57.34	50.78	50.51	49.51	51.20	50.28
4	54.09	58.89	58.12	64.75	48.99	53.06	61.65	**65.71**	50.51	53.31	54.79	51.76
5	60.15	72.84	**79.46**	78.72	49.74	59.15	72.59	69.53	50.51	51.28	67.27	50.02
6	88.60	77.70	**89.37**	87.06	49.49	58.54	72.53	59.39	50.51	50.76	65.46	50.26
7	51.30	**56.88**	54.32	56.56	49.49	54.50	53.60	53.77	51.02	49.5	54.32	50.97
8	63.21	61.38	**64.45**	63.94	48.98	57.36	57.91	55.40	50.51	51.00	51.75	49.23
9	**96.70**	84.25	96.43	96.67	49.49	87.29	92.15	91.64	50.51	51.02	79.92	50.51
10	47.71	51.78	47.76	43.38	48.24	50.00	51.33	**53.56**	50.51	49.23	50.73	49.49
AVE	63.2	68.55	**72.35**	72.02	49.39	63.06	68.96	65.21	50.56	50.72	61.50	50.51
SD	20.43	11.14	15.50	15.69	8.50	8.15	10.42	8.70	0.16	1.05	10.10	0.82

Table 6.3 Pairwise t-test between results of the proposed DTW features and the time-domain feature sets in classifying speech versus non-speech using mouse click separated data;✓ means significant, × means not significant. The values inside the parenthesis are p values

Compared Features	Classification Algorithms			
	SVM	NB	RF	LDA
DTW and MaxCC	×	✓(0.02)	×	✓(0.05)
DTW and Statistics of EEG	×	✓(0.001)	✓(0.003)	✓(0.005)

6.6.2 Classifying the five imagined words for mouse click separated data using time-domain features

Table 6.4 shows the average classification accuracy for classifying the five words using the TD feature sets and the four classifiers. As shown, the best feature was the proposed DTW using LDA classifier (52.50%). The second-best set of features was MaxCC. For MaxCC and the statistics of EEG, RF yielded the highest average accuracy. In general, LDA and RF are good classifiers for EEG data, and they have been used in several previous studies. The pairwise t-tests showed that, for all the classifiers, the proposed DTW features significantly outperformed the other feature sets (Table 6.5).

Toward understanding merits of the proposed DTW-based features

As previously explained, each feature was defined as the total distance between each pair of EEG trials after warping them using DTW. As seen in Table 6.4, subject 5 had the highest classification accuracy when classifying five words using the proposed DTW. Calculating the confusion matrix for the LDA classification results of this subject, the word, *"select"*, had the best classification accuracy (Table 6.6). Figure 6.3 illustrates the distribution of the distances between the trials of *"select"* and the other words for subject 5. The DTW distances were significantly smaller between the trials of *"select"* than between the trials of the other words (Figure 6.3). Some small overlap was noticed when comparing the distances of the trials of "select" and the trials of *"up"*.

Table 6.4 Average 10-fold cross-validation results (%) of classifying the five imagined words for mouse click separated data using three time-domain feature sets across four classifiers; the best result for every subject is in bold

Subject	DTW				MaxCC				Statistics of EEG			
	SVM	NB	RF	LDA	SVM	NB	RF	LDA	SVM	NB	RF	LDA
1	39.33	38.00	57.00	**58.00**	20.67	33.67	43.00	35.67	32.33	33.00	34.33	21.33
2	36.67	31.3	**44.3**	42.67	19.67	28.00	32.67	25.33	22.67	20.00	24.67	20.00
3	49.00	32.67	46.00	**49.67**	20.67	25.33	21.67	36.33	22.67	25.33	31.67	20.67
4	47.67	28.00	50.33	**51.00**	20.67	30.67	40.00	33.67	24.00	20.33	38.00	20.67
5	47.33	60.67	**75.00**	**75.00**	24.67	44.33	59.67	50.00	41.00	21.00	27.33	31.67
6	32.00	27.67	30.00	**35.67**	20.00	22.00	21.00	15.67	20.00	20.00	28.00	21.00
7	17.00	33.67	54.00	**57.67**	20.67	30.67	37.67	42.33	22.33	20.33	23.67	21.33
8	30.33	29.00	41.00	**50.00**	22.00	33.00	30.33	32.00	21.667	20.33	34.33	22.33
9	47.00	32.00	44.33	**49.33**	20.33	30.67	33.00	34.33	20.00	20.00	22.67	21.00
10	37.33	42.67	46.33	**56.00**	20.00	31.67	38.67	35.00	21.00	19.00	33.67	17.67
AVE	38.37	35.57	48.83	**52.50**	20.93	31	35.77	33.78	24.77	22.17	29.70	21.77
SD	10.11	9.64	11.76	10.37	1.41	5.88	11.09	1.16	6.37	4.40	4.90	3.50

Table 6.5 Pairwise t-test between results of the proposed DTW features and time-domain feature sets in classifying the five imagined words(mouse click separated data);✓ means significant, × means not significant. The values inside the parenthesis are p values

Compared Features	Classification Algorithms			
	SVM	NB	RF	LDA
DTW and MaxCC	✓(0.0003)	✓(0.00001)	✓(0.00001)	✓(0.00000005)
DTW and Statistics of EEG	✓(0.002)	✓(0.002)	✓(0.001)	✓(0.0000009)

This issue also can be seen in the confusion matrix shown in Table 6.6 where *"select"* was misclassified as *"up"* in 6.67% of the trials. Similarly, *"up"* was misclassified as *"select"* in 13.33% of the trials.

Table 6.6 Confusion matrix for classifying the five imagined words using the proposed DTW-based features and LDA classifier for subject 5

Word	Left	Right	Up	Down	Select
Left	88.33	5.00	1.67	0.00	5.00
Right	6.67	45.00	18.33	8.33	21.67
Up	0.00	3.33	80.00	3.33	13.33
Down	1.67	0.00	11.67	68.33	18.33
Select	0.00	0.00	6.67	0.00	93.33

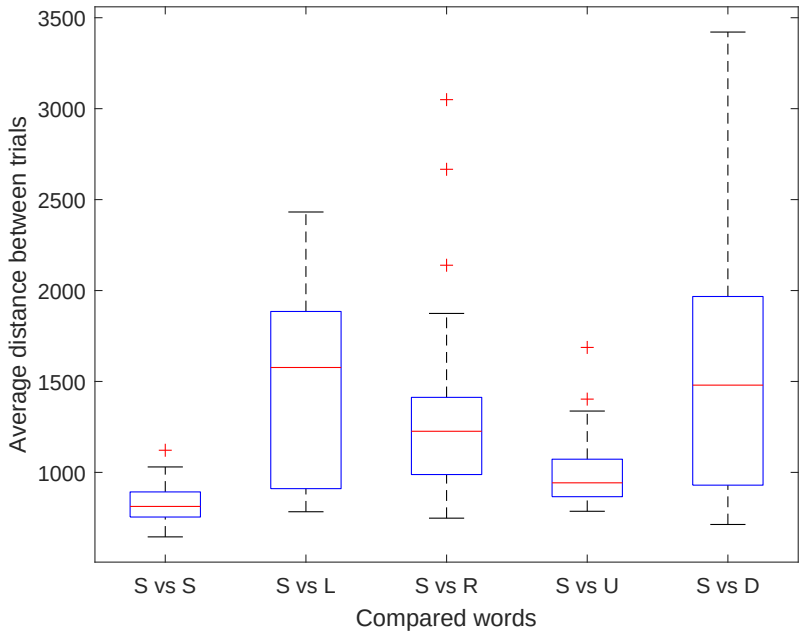

Fig. 6.3 Distribution of average DTW distances between trials of the word "select" for subject 5 and distances between the word "select" and other words. (The letters represent the first letter from each word).

6.6.3 Classifying between imagined speech and silence for fixed time separated data using time-domain features

For the fixed-time separated EEG trials, the classification between speech and non-speech using the proposed DTW framework was compared with the TD features. The examined time from each class was two seconds. For speech imagination, since the trial length was four seconds, only the first two seconds were used to compare

with the non-speech time. Table 6.7 presents the average classification of imagined speech and non-speech using the compared TD features and the four classifiers. The proposed DTW feature extraction, using LDA as the classifier, provided the best average classification accuracy in comparison to the other feature sets and classifiers. This result is statistically significant, except for the MaxCC features using the LDA classifier as in Table 6.8.

Classifying imagined speech versus non-speech using the proposed DTW and LDA classifier for fixed-time separated had less average classification accuracy than mouse click separated data (Table 6.2). This finding is discussed in more details in Section 6.7.

Table 6.7 Average 10-fold cross-validation results (%) of classifying unspoken speech versus silence fixed time separated data using three time-domain feature sets across four classifiers; the best result for every subject is in bold

Subject	DTW				MaxCC				Statistics of EEG			
	SVM	NB	RF	LDA	SVM	NB	RF	LDA	SVM	NB	RF	LDA
1	84.22	73.18	85.71	90.23	48.88	77.72	87.22	**91.48**	49.87	50.13	76.44	49.87
2	50.86	50.61	47.88	50.38	50.87	49.87	48.37	50.86	49.87	48.38	51.38	49.12
3	54.15	45.61	46.62	**58.39**	48.11	50.37	55.89	53.39	49.87	48.62	49.36	49.87
4	50.86	58.13	59.64	60.64	50.38	60.65	63.18	**64.39**	49.87	50.13	56.13	50.13
5	57.04	**60.15**	59.80	59.54	52.04	56.72	58.60	56.39	49.83	49.84	53.27	49.84
6	55.38	53.61	56.39	**65.16**	50.13	52.60	59.42	55.64	49.87	50.38	51.63	45.59
7	57.63	55.88	59.39	69.92	50.13	59.66	62.67	**71.43**	49.87	49.37	50.61	49.62
8	52.63	50.87	45.36	49.17	49.63	52.89	52.40	51.11	49.87	50.88	50.11	49.37
9	56.40	52.14	52.88	58.14	50.37	51.63	54.38	**61.41**	49.87	50.38	50.89	49.88
10	59.66	49.87	61.14	**67.65**	50.13	59.38	61.90	58.87	49.87	49.62	53.64	48.87
AVE	57.88	55.00	57.48	**62.92**	56.16	55.01	54.89	56.07	49.87	49.77	54.35	49.21
SD	9.67	7.45	11.49	11.57	8.60	7.82	7.81	8.91	0.01	0.79	8.01	1.33

Table 6.8 Pairwise t-test between results of the proposed DTW features and the time-domain feature sets in classifying speech versus non-speech (fixed time separated data);✓ means significant, × means not significant. The values inside the parenthesis are p values

Compared Features	Classification Algorithms			
	SVM	NB	RF	LDA
DTW and MaxCC	✓(0.03)	✓(0.10)	✓(0.0168)	×
DTW and Statistics	✓(0.02)	✓(0.05)	×	✓(0.005)

Table 6.9 Average 10-fold cross-validation results (%) of classifying the five imagined words for fixed time separated data using three feature-extraction methods across four classifiers; the best result for every subject is in bold

Subject	DTW				MaxCC				Statistics of EEG			
	SVM	NB	RF	LDA	SVM	NB	RF	LDA	SVM	NB	RF	LDA
1	47.73	44.71	57.79	**71.90**	23.67	31.60	26.17	40.19	30.65	21.60	40.69	21.60
2	59.23	30.23	42.67	**75.38**	21.63	23.69	29.63	38.71	43.71	30.19	60.29	20.60
3	41.65	40.25	44.25	**65.33**	16.10	33.08	34.66	29.13	27.10	30.67	58.79	21.69
4	71.88	45.73	73.92	**93.46**	20.60	23.13	32.73	35.69	20.10	22.60	52.69	26.10
5	69.77	47.17	69.70	**87.34**	24.51	29.54	38.91	52.11	49.01	28.29	48.32	27.66
6	41.65	46.69	56.25	**76.85**	20.10	30.63	26.13	34.67	20.60	19.60	63.88	20.10
7	61.15	68.29	83.92	**94.46**	24.60	59.79	70.46	77.42	20.10	20.08	21.58	29.21
8	41.75	40.73	59.77	**72.31**	23.06	35.73	50.75	40.79	46.75	20.63	35.73	18.56
9	81.44	54.31	82.96	**95.48**	22.65	45.21	58.27	57.71	33.69	39.71	74.29	20.10
10	30.60	34.71	54.29	**54.77**	22.10	33.19	39.21	34.19	27.65	21.10	51.31	15.58
AVE	54.68	45.28	62.55	**78.73**	21.90	34.56	40.69	44.06	31.94	25.45	50.76	22.12
SD	14.11	9.87	14.32	10.74	2.53	10.81	14.70	14.10	6.37	4.40	4.90	3.50

6.6.4 Classifying the five imagined words for fixed time separated data using time-domain features

For each subject, one block (40 trials) from each word was recorded where the word imagination was 4 second (see Chapter 5, section 5.1.3).

Table 6.9 shows the classification of the five imagined words using the proposed DTW and two TD feature sets using four classifiers. For all subjects, the proposed DTW features using the LDA classifier significantly outperformed the other feature sets. Table 6.10 presents pairwise t-tests results of using the proposed DTW in comparison to the other feature sets.

Table 6.10 Pairwise t-test between results of the proposed DTW features and the other feature sets in classifying the five imagined words (fixed time separated data);✓ means significant, × means not significant. The values inside the parenthesis are p values

Compared Features	Classification Algorithms			
	SVM	NB	RF	LDA
DTW and MaxCC	✓(0.0001)	✓(0.0005)	✓(0.0001)	✓(0.000004)
DTW and Statistics	✓(0.003)	✓(0.0007)	×	✓(0.00000005)

The results in Table 6.9 were then compared with the classification of imagined words using mouse-click separated data. Table 6.11 presents the classification of the five words using 40 trials for each word for mouse click separated data. In contrast to classifying speech and non-speech results, the fixed-time separated data results were significantly higher than the mouse-click separated data results (see sections 6.6.1 and 6.6.3). This could be due to several factors, including the imagination length, the overlap between the imagination task and the motor execution task using mouse clicks. More discussion of the results is in Section 6.7.

Table 6.11 10-fold cross-validation classification accuracy (%) to classify the five imagined words using the proposed DTW and four classifiers (mouse click separated data);using 40 EEG trials for each word; t-tests compare the average classification of mouse click separated words with the average classification of fixed time separated words for each classifier

Subject	SVM	NB	RF	LDA
1	49.37	43.68	58.82	70.87
2	47.18	38.68	49.16	53.24
3	55.74	36.18	40.16	58.37
4	57.26	30.05	43.08	65.32
5	84.45	76.37	81.39	87.50
6	41.66	27.61	29.08	46.68
7	35.68	26.08	50.79	49.82
8	44.68	29.63	40.68	50.24
9	65.82	38.24	47.79	65.32
10	63.34	49.18	53.21	69.82
AVE	54.52	39.57	49.42	61.72
SD	13.83	14.49	13.86	12.25
T-test	×	×	✓(0.04)	✓(0.015)

Toward understanding merits of the proposed DTW-based features

In the proposed DTW features, the variation in the distances between the two warped signals represents the features distribution. Figure 6.4 shows the distribution of distances for the word *"up"* in comparison to other words for subject 9. Subject 9 had the best classification accuracy (95.48%) using LDA. Moreover, *"up"* has the best

accuracy of the five words (100%), as can be seen in the confusion matrix in Table 6.12. The distances between trials for word "up" did not completely overlap with the distances with the other words, as in Figure 6.4.

Subject 10 had the lowest classification accuracy (54.77%) using LDA. The word *"select"* had the lowest classification (25.00%) among the five words (Table 6.12). As seen in Figure 6.5, an overlap exists when comparing the distances of the trials of the word *"select"* and the trials for the other words. Most of the misclassifications were between *"select"* and *"right"*. This was because the average distance between these two words was very small and overlapped with the distances of the word *"select"* trials.

Table 6.12 Confusion matrix for classifying the five imagined words using DTW-based features and LDA classifier for subject 9

Word	**Left**	**Right**	**Up**	**Down**	**Select**
Left	100.00	0.00	0.00	0.00	0.00
Right	0.00	100.00	0.00	0.00	0.00
Up	0.00	0.00	100.00	0.00	0.00
Down	0.00	5.00	7.50	85.00	2.50
Select	0.00	2.50	0.00	2.50	95.00

Table 6.13 Confusion matrix for classifying the five imagined words using DTW-based features and LDA classifier for subject 10

Word	**Left**	**Right**	**Up**	**Down**	**Select**
Left	50.00	37.50	2.50	0.00	10.00
Right	2.50	90.00	5.00	0.00	2.50
Up	5.00	32.50	60.00	2.50	0.00
Down	2.50	2.50	32.50	55.00	7.50
Select	5.00	47.50	5.00	12.50	30.00

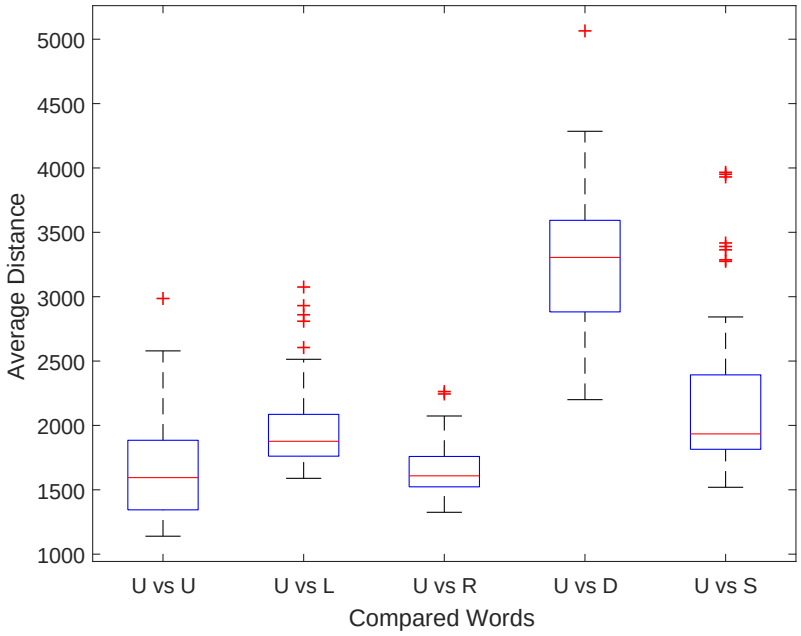

Fig. 6.4 Distribution of average DTW distances between trials of the word "up" for subject 9 and DTW distances between the word "select" and other words. (The letters represent the first letter from each word).

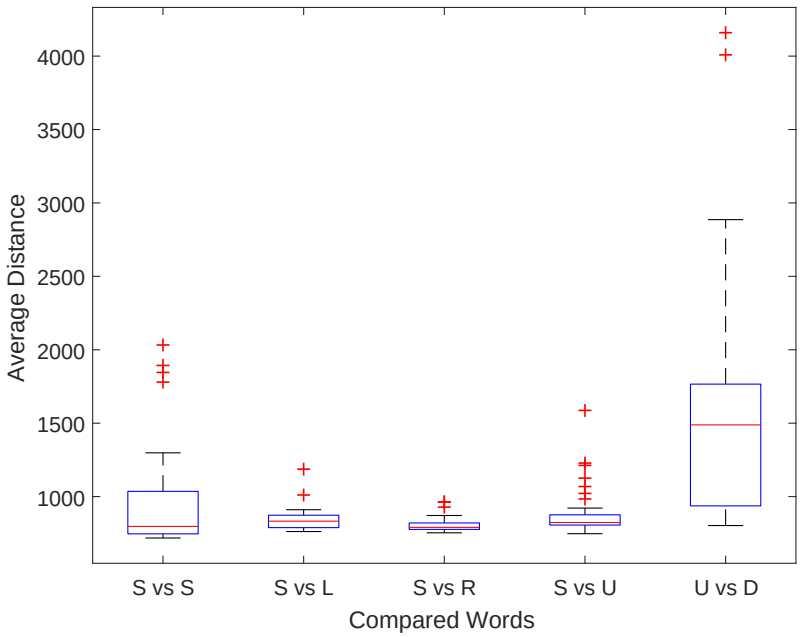

Fig. 6.5 Distribution of average DTW distances between trials of the word "select" for subject 10 and DTW distances between the word "select" and other words. (The letters represent the first letter from each word).

6.6.5 Classifying the five imagined words for mouse click separated data using DDTW feature sets

Section 6.4.4 explains derivative dynamic time warping (DDTW), which was a proposed modification of classical DTW. Kumagai et al. (2017) claimed that DDTW solves some problems during signal warping. In the present study, DDTW was used to classify the five imagined words using mouse-click separated data applying the steps used for the proposed DTW. Table 6.14 lists the average classification accuracy across subjects of classifying the five imagined words and the pair-wise t-test results in comparison to the proposed DTW framework.

The results show that the proposed DTW statistically outperformed DDTW with the SVM and LDA classifiers. However, the difference is not significant in comparison to the RF and NB classifiers. This finding could be justified based on the differences between DTW and DDTW and the technical differences between the classifiers. DDTW works on the derivative of the signal, which amplifies the noise more than the original signal (Jauberteau and Jauberteau, 2009). Moreover, LDA and SVM are known to be the best classifiers for brain computer interface studies. However, their results can be easily affected by noise (Müller et al., 2004). The NB classifier is similar to the LDA classifier in terms of assuming Gaussian distribution of the data. However, NB assumes the features are independent (Misaki et al., 2010). Consequently, the results are stable when using either the derivative or the original signal. The RF classification algorithm performs the decision based on a subset of the features. Typically, this makes it successful and stable with noise and a small training size (Lotte et al., 2018).

Table 6.14 Average classifications results (across subjects) (%) of classifying the five imagined words for mouse click separated data using DDTW feature sets; pairwise t-tests to compare the classification accuracies resulted from using the proposed DTW and DDTW for each classifier;✓ means significant, × means not significant. The values inside the parenthesis are p values

Method	SVM	NB	RF	LDA
DTW	38.37	35.57	48.83	**52.50**
DDTW	21.93	34.23	51.00	43.77
T-test	✓(0.007)	×	×	✓(0.0001)

6.6.6 Classifying the five imagined words (mouse click separated data, fixed time frame separated data) using all channels

Section 6.4.4 explained the differences between using EEG data from all channels, (the approach for audio-speech recognition in DTW), and using the proposed framework. Here, the two methods are compared for both mouse-click separated data and fixed-time separated data. The results are shown in Table 6.15 and Table 6.16. As seen, the proposed DTW feature extraction outperformed the use of all channels as single inputs to DTW. This could be due to differences between the recognition of brain and speech signals. The audio-speech signal only contains the speech event that occurs in a serial process, during which any delay in one part affects the rest. Consequently, the inputs to the frequency channels are compatible in time, and the DTW computation from all channels can be performed in one step. However, the brain does not work in serial steps (Vaadia and Birbaumer, 2009). The response to each sensory input comes through several computational complex internal models. This makes it difficult to expect the time between the inputs to EEG channels to be compatible.

Table 6.15 Average classification accuracy (across subjects) (%) to classify the five imagined words using the proposed DTW and DTW using all channels (mouse click separated data);pairwise t-tests to compare the classification accuracies resulted from using the proposed DTW and DTW using all channels for each classifier; ✓means significant, × means not significant. The values inside the parenthesis are p values

Method	SVM	NB	RF	LDA
Proposed DTW	38.37	35.57	48.83	**52.50**
DTW using all channels	21.91	33.13	43.34	42.89
T-test	✓(0.0004)	✓(0.007)	✓(0.01)	✓(0.00002)

Table 6.16 10-fold cross-validation classification accuracy (%) to classify the five imagined words using the proposed DTW and DTW using all channels (Fixed time separated data); pairwise t-tests to compare the classification accuracies resulted from using the proposed DTW and DTW using all channels for each classifier; ✓means significant, × means not significant. The values inside the parenthesis are p values

Method	SVM	NB	RF	LDA
Proposed DTW	54.68	45.28	62.55	**78.73**
DTW using all channels	30.18	31.81	36.60	45.93
T-test	✓(0.0001)	✓(0.0004)	✓(0.000005)	✓(0.000001)

6.6.7 Improving the DTW-based framework by removing outliers for classifying the five imagined words using mouse click separated data

The DTW-based framework could be improved by removing outlier-training trials. A trial is an outlier when the distance between it and the rest of the trials from the same class is closer than the distance between it and the trials from the other classes. This closer distance could be due to several factors. It could be affected by the length similarity between the words, which is an important aspect of speech rate. Artefact contamination could also make the distance unrealistic for some trials. The third factor is learning, since the subjects were instructed to be consistent during the recording of each word. Over time, the learning factor would be an important factor in classification accuracy. Fatigue may also have contributed to lack of consistency between the trials.

Figure 6.6 shows the classification accuracies after removing the outlier training trials. For the SVM classifier, an average improvement of 10.33% was seen after removing 34 trials from each word. The accuracies vary slightly (both increasing and decreasing) for the RF and LDA classifiers. For NB, removing four trials yielded a 3.00% improvement; afterwards, the accuracy is mostly stable. In comparison to LDA and NB, SVM and RF work based on a selected subset of the features. However, SVM works based on identifying the boundaries according to the distance between the features in the same class. In contrast, RF works based on the probability of belonging to the same class (Lotte et al., 2007). Thus, reducing the number of training results (number of features) helps improve the features optimization process for SVM.

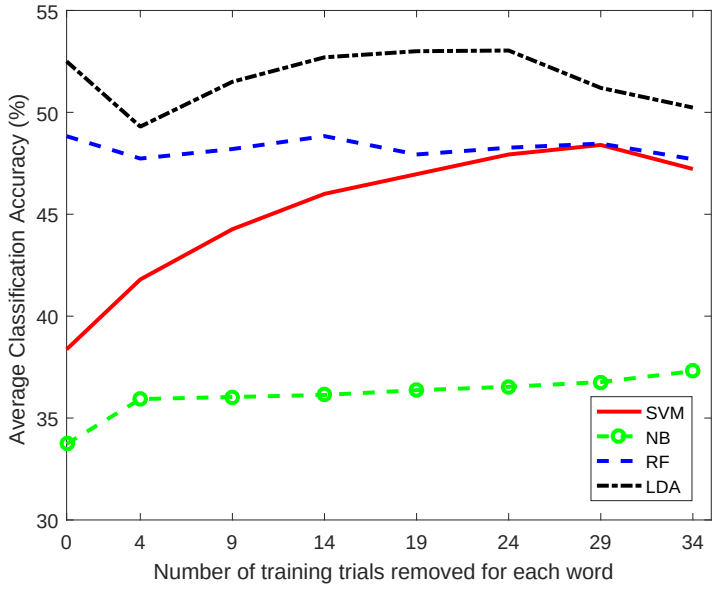

Fig. 6.6 Average classification accuracy across all subjects after removing outliers from training trials.

6.6.8 Classifying between speech and silence (mouse click separated data, fixed time frame separated data) using time-frequency feature sets

Table 6.17 compares the 10-fold classification accuracies between the imagined words versus silence for the mouse-click separated data using the proposed DTW. It also presents the accuracies for the two time-frequency feature sets based on DWT coefficients: RWE-DWT and statistics-DWT. Only the results from LDA and RF were compared. As they were the best classifiers based on the results presented in Chapter 5 (see Section 5.5). Moreover, LDA provided the best classification of the proposed DTW features in the experiments previously presented in this chapter. The proposed DTW using RF provided the best average classification accuracy compared to the two DWT-based feature sets. The statistical significance of the results are shown in Table 6.18.

Table 6.17 Average 10-fold cross-validation results (%) of classifying unspoken speech versus silence for mouse click separated data using the proposed DTW and two time-frequency methods across two classifiers; the best result for each subject is in bold

Subject	DTW		DWT-RWE		Statistics-DWT	
	RF	LDA	RF	LDA	RF	LDA
1	81.24	**81.49**	71.54	69.83	69.79	75.89
2	93.64	**95.91**	83.49	75.70	76.45	90.88
3	**58.66**	51.75	52.49	50.79	52.05	56.81
4	58.12	64.75	53.01	52.32	58.38	51.99
5	**79.46**	78.72	66.00	61.19	63.09	78.92
6	**89.37**	87.06	69.28	67.24	67.03	82.76
7	54.32	56.56	54.34	54.92	51.25	51.25
8	**64.45**	63.94	57.16	52.48	53.60	60.64
9	**96.43**	96.69	92.15	91.65	73.07	89.60
10	47.76	43.38	49.26	48.24	47.93	48.96
AVE	**72.35**	72.02	63.14	60.15	61.27	68.77
SD	15.50	15.69	10.42	8.70	8.89	15.00

Table 6.18 Pairwise t-test between results of the proposed DTW features and time-frequency feature sets in classifying speech versus non-speech using mouse click separated data;✓ means significant, × means not significant. The values inside the parenthesis are p values

Compared Features	Classification Algorithms	
	RF	LDA
DTW and RWE-DWT	✓(0.004)	✓(0.004)
DTW and DWT-Statistics	✓(0.003)	×

Table 6.19 presents the 10-fold classification accuracies between the imagined words and silence using fixed-time separated data using the proposed DTW, and the two time-frequency feature sets. Only the first 2 seconds from the speech imagination trials was used. The results show that the average classification accuracy using the proposed DTW framework significantly outperformed the other features using the LDA classifier (Table 6.20). In conclusion, the proposed DTW feature extraction framework provides the best feature sets to classify imagined speech versus silence (Table 6.17 and Table 6.19).

Table 6.19 Average 10-fold cross-validation results (%) of classifying unspoken speech versus silence for fixed time separated data using DTW and two time-frequency methods feature sets across two classifiers; the best result for every subject is in bold

Subject	DTW		DWT-RWE		DWT-Statistics	
	RF	LDA	RF	LDA	RF	LDA
1	85.71	**90.23**	74.46	78.21	79.22	77.45
2	47.88	50.38	48.62	**50.86**	48.13	46.37
3	46.62	**58.39**	49.38	56.38	47.37	53.38
4	59.64	**60.64**	58.67	56.15	54.64	53.38
5	**59.80**	59.54	48.61	49.26	57.05	52.67
6	56.39	65.16	50.13	51.63	57.88	51.64
7	59.39	**69.92**	59.40	58.38	58.65	59.14
8	52.88	**58.14**	55.61	55.90	52.62	56.13
9	52.88	58.14	55.61	55.89	**59.13**	54.64
10	**61.14**	67.65	57.14	55.88	59.13	54.64
AVE	57.48	**62.92**	54.89	56.08	55.80	55.78
SD	11.49	11.57	7.81	8.91	9.36	8.32

Table 6.20 Pairwise t-test between results of the proposed DTW features and time-frequency and CSP feature sets and the proposed DTW in classifying speech versus non-speech using mouse click separated data;✓ means significant, × means not significant. The values inside the parenthesis are p values

Compared Features	Classification Algorithms	
	RF	LDA
DTW and RWE-DWT	✓(0.004)	✓(0.004)
DTW and DWT-Statistics	×	✓(0.0021)

6.6.9 Classifying between the imagined words (mouse click separated data, fixed time frame separated data) using time-frequency feature sets

The classification accuracies of the proposed DTW and DWT-based features (RWE-DWT and statistics-DWT) were compared for mouse-click separated data and fixed-time separated data in classifying the five imagined words. The average classification accuracy using the statistics-DWT features with the RF classifier statistically outperformed the proposed DTW features (Tables 6.21 and 6.23). The proposed DTW worked better with the LDA classifier, but it is less effective than the statistics-DWT using RF (Table 6.22 and Table 6.24).

Table 6.21 Average 10-fold cross-validation results (%) of classifying imagined words for mouse click separated data using DTW and two time-frequency methods across two classifiers; the best result for every subject is in bold

Subject	DTW		DWT-RWE		Statistics-DWT	
	RF	LDA	RF	LDA	RF	LDA
1	57.00	58.00	46.00	48.00	**66.67**	48.33
2	44.33	42.67	26.33	33.00	**55.33**	45.00
3	46.00	49.67	26.67	25.33	**58.67**	46.33
4	50.33	51.00	33.00	37.67	**73.67**	52.00
5	75.00	75.00	50.33	45.67	**83.33**	57.67
6	30.00	35.67	24.33	28.33	**47.67**	36.00
7	54.00	57.67	44.33	46.00	**71.67**	55.67
8	41.00	50.00	28.67	35.33	**53.67**	44.00
9	44.33	49.33	26.00	26.33	**55.67**	40.33
10	46.33	56.00	32.67	33.00	**65.33**	46.67
AVE	48.83	52.50	33.83	35.87	**63.17**	47.2
SD	11.19	9.10	9.05	7.90	10.37	6.28

Table 6.22 Pairwise t-test between results of the proposed DTW features and time-frequency feature sets in classifying the imagined words using mouse click separated data;✓ means significant, × means not significant. The values inside the parenthesis are p values

Compared Features	Classification Algorithms	
	RF	LDA
DTW and RWE-DWT	✓(0.00001)	✓(0.00006)
DTW and DWT-Statistics	✓(0.000005)	✓(0.02)

Table 6.23 Average 10-fold cross-validation results (%) of classifying imagined words for fixed time separated data using DTW and three time-frequency methods across two classifiers; the best result for every subject is in bold

Subject	DTW		DWT-RWE		Statistics-DWT	
	RF	LDA	RF	LDA	RF	LDA
1	57.79	71.90	68.42	72.88	**88.87**	55.76
2	42.67	75.38	57.29	61.31	**81.47**	74.87
3	44.25	65.33	49.73	41.21	**85.92**	60.76
4	73.92	93.46	70.83	68.33	**97.00**	70.87
5	69.70	87.34	45.92	49.11	**89.92**	38.96
6	56.25	76.85	56.23	51.31	**77.37**	53.66
7	83.92	94.46	67.38	70.81	**92.92**	78.87
8	59.77	72.31	56.29	48.29	**83.37**	52.82
9	82.96	95.48	70.29	65.31	**97.97**	76.89
10	54.29	54.77	37.67	35.71	**71.87**	38.18
AVE	62.55	78.73	58.00	56.43	**86.67**	60.16
SD	13.86	12.94	10.71	12.36	7.96	14.18

Table 6.24 Pairwise t-test between results of the proposed DTW features and time-frequency feature sets in classifying the imagined words using fixed time separated data;✓ means significant, × means not significant. The values inside the parenthesis are p values

Compared Features	Classification Algorithms	
	RF	LDA
DTW and RWE-DWT	×	✓(0.00007)
DTW and DWT-Statistics	✓(0.00004)	✓(0.0013)

6.6.10 Classifying between speech and silence (fixed time frame separated data) using CSP feature sets

Common spatial patterns is a technique to maximise the variance between EEG signals. CSP was mainly designed to discriminate between two conditions and it requires the two signals to be in the same length. Section 4.2.6 provided more information about CSP. CSP was successfully applied in motor imagination classification, in some speech imagination studies (in Chapter 3) and in the experiment explained in Chapter 4. Table 6.25 presents the comparison between the proposed DTW and CSP in classifying unspoken speech versus silence. The proposed DTW significantly outperformed CSP using LDA classifier.

Table 6.25 Average 10-fold cross-validation results (%) of classifying unspoken speech versus silence for fixed time separated data using DTW and CSP features across two classifiers; the best result for every subject is in bold; t-tests compare the proposed DTW and CSP for each classifier;✓ means significant, × means not significant. The values inside the parenthesis are p values

Subject	DTW		CSP	
	RF	LDA	RF	LDA
1	85.71	**90.23**	81.75	81.75
2	47.88	**50.38**	47.75	47.5
3	46.62	**58.39**	50.5	50.25
4	59.64	60.64	58.00	**61.75**
5	**59.80**	59.54	50.31	55.94
6	56.39	65.16	52.25	58.5
7	59.39	69.92	64.00	**72.25**
8	52.88	**58.14**	47.25	44.25
9	52.88	58.14	55.75	53.50
10	**61.14**	67.65	55.75	53.50
AVE	57.48	**62.92**	55.38	57.52
SD	11.49	11.57	10.77	11.63
T-test	RF:	×	LDA:	✓(0.018)

6.7 Discussion and conclusions

This chapter presents the first use of DTW for improving the classification of imagined speech via EEG signals. The classified patterns consisted of five imagined words: *"left"*, *"right"*, *"up"*, *"down"* and *"select"*. The data acquisition from 10 subjects was performed using a wireless EEG device with only 14 channels. Each word was recorded using two different trials separation methods: mouse-click separation and fixed-time separation. The proposed framework based on DTW was evaluated in comparison to TD features and time-frequency features, CSP, as well as some modifications to the proposed framework. These comparisons were performed using four classification algorithms: support vector machine (SVM), naive Bayes (NB), random forests (RF) and linear discriminant analysis (LDA). The experimental assessments involved discriminating between speech and non-speech and discriminating between classification of the five imagined words.

In summary, the proposed DTW feature extraction framework outperformed the TD features in the experiments 1 to 4. In experiment 5, the proposed DTW framework for the imagined words classification using LDA outperformed DDTW. In experiment 6, the proposed DTW feature extraction framework outperformed the classical DTW used in audio-speech recognition. In experiment 8, the proposed DTW framework outperformed time-frequency features in classifying imagined speech versus silence. In contrast, the average classification accuracy for classifying the five imagined words was higher with the statistics-DWT than the proposed DTW.

In comparison to the other TD features, DTW matches EEG signals without having to average or remove any parts from the signal. Moreover, this mapping is not one-to-one in which the variations between the start and the end are considered. This makes the resulting distance reflect the level of similarity between the compared EEG patterns. In comparison to RWE-DWT, calculating the relative wavelet energy from the DWT coefficients, as in (González-Castañeda et al., 2017), reduced the temporal information amount. It also reduced the effectiveness of the DWT coefficients, as discussed in (Yohanes et al., 2012). Comparably, statistics-DWT (RMS and SD on

the DWT coefficients) reflects the important role of frequency information in the classification of imagined words.

Three modifications were applied to the proposed DTW framework. One modification was using a derivative DDTW. Another modification was removing outlier training trials. The third modification was using EEG data from all channels as one input in the computation of DTW distance. In DDTW (experiment 5), the computation of the derivative of the signal amplified the noise. Since EEG is known to have a very noisy signal, this decreases the classification accuracy for the LDA and SVM classifiers. The removal of the outlier trials was suggested to enhance the performance of the proposed framework (experiment 7). The enhancement was clear in the SVM classifier results. However, it was not significant in the other classifiers. The last proposed modification was to use EEG from all the channels as a signal input to the DTW. The results showed that the proposed framework performed significantly better when the complexity of the generated brain signals cannot be compared with speech signals.

In terms of the differences in the results based on the trials separation method, the mouse-click separated data had better average classification accuracy for speech and non-speech than the fixed-time separated data. However, the fixed-time separated data provided better classification of the five imagined words. This could be due to two factors. First, in the classification of speech and non-speech, the difference between the two classes also includes the differences between the imagined words. In terms of results, the variation in the imagination length of speech and non-speech would help improve the classification accuracy. Second, for word classification, increases in the length of imagination time provides extra information (padding) to the signal (see Chapter 5, Section 5.3.2). This helps in distinguishing among the five words. Moreover, the mouse-click separated data included extra patterns, beyond word imagination patterns. These patterns were intended to perform the mouse click at the beginning and end of the task. In addition, muscle movements occurred during the clicks. These movements were very small; the usual time needed for adults to perform mouse clicks is 100 ms (Komandur et al., 2008).

6.8 Summary

The recognition of unspoken speech could be the most intuitive type of brain-computer interface for people with severe speech disabilities. Consequently, researchers are increasingly interested in classifying different types of unspoken speech from EEG signals. However, the time variations of imagination reflected in EEG signals have not been considered in previous studies. These variations, caused by differences in the starting time and duration of the imagined words, could have a detrimental effect on classification accuracy. In this chapter, for the first time, these temporal variations were investigated and minimised using a DTW-based framework. In this technique, the distances between the imagined words after warping by DTW are used as classification features. The proposed DTW framework was evaluated using EEG data collected from 10 subjects who imagined five different words. The evaluation involved discriminating between imagined speech and silence. It also involved discriminating between five imagined words from two data sets based on different trial separation methods (mouse-click separation and fixed-time separation).

10 experiments were conducted. These compared the classification accuracy results from the proposed DTW features and state-of-the-art features, and three modifications to the proposed framework (Table 6.1). The results show that the DTW-based framework outperformed all the discussed state of the-art feature extraction algorithms in classifying imagined speech versus non-speech. The proposed framework also outperformed the TD features in classifying the five imagined words. The justifications of these results were described in Section 6.7.

Chapter 7

Conclusions

This chapter summarises the findings from the research reported in this thesis with respect to the objectives listed in Chapter 1. It then outlines its contributions to the imagined speech research domain. Finally, it offers directions for future work inspired by the results of the experiment performed so far.

7.1 Reviewing thesis scope and main findings

The research described in this thesis was motivated by the need to understand and alleviate several limitations in the recognition of imagined speech using EEG signals; it had three primary objectives:

1. Improving the discrimination between speech and non-speech.

2. Optimising a computational model to improve the classification between the imagined words by examining several temporal variations in the recognition model. This involved using EEG pattern separation methods, establishing different time intervals and examining the effect of word length in the recognition.

3. Improving imagined speech recognition by reducing the variations between EEG trials using the dynamic time warping (DTW) algorithm.

Several steps (as presented in the related chapters) were used to achieve these objectives:

The first stage in this research was to identify the major challenges and limitations in recognising imagined speech research studies. Chapter 3 presents a literature survey on the studies in the context of imagined speech recognition using brain–computer interface technologies. It concentrated on studies of EEG signals (the technology of interest in this thesis). The main conclusions from this chapter can be summarised as follows:

- The research on imagined speech recognition using EEG signals is a relatively new research domain. The first study to recognise imagined words was conducted by Wester (2006). The research studies conducted from 2006–2016 (before the beginning of this research) had limited results due to a lack of available datasets. After this, interest in the research domain increased, and several methods and results emerged.
- Compared to other applications for EEG (motor imagination), prior research had inconsistencies in their experimental design and data collection methods.
- Most of the studies focused on recognising speech stimuli based on phonological differences.
- The studies had a limited understanding of the recognition of imagined speech compared to non-speech task.
- Similar to other applications of BCI, there was a limited understanding of the contributions of temporal information to improving BCI recognition.

In Chapter 4, the first objective was achieved in terms of classifying imagined speech versus non-speech tasks. EEG data were collected from nine subjects during the imagination of semantically varying words. The literature presented evidence of the impact of word semantics on brain signals. The non-speech tasks asked participants to concentrate on visualised stimuli on a screen (the presentation of '+' and the presentation of the word) and silence time. The data analysis involved examining

time domain (statistics of EEG) and spatio-spectral features (filter bank common spatial patterns) at different time intervals using different classifiers. The classification accuracies were examined for each word and for groups of words compared to non-speech tasks. The results showed differences in classification accuracy for different subjects and different features.

To achieve the second objective, Chapter 5 described the examination of important temporal experiment parameters in designing imagined speech recognition experiments. EEG data related to the imagination of five words were collected from 10 subjects. For each subject, each word was recorded with two different trial-separation methods: mouse-click separation and specified time frame separation. The experimental aspects examined in the study showed that the specification of long time frames provided distinguishable EEG patterns. The increase in training size also improved classification accuracy, although these improvements lessened after a certain training size. In addition, if the recording was performed at different times in the session, then the training size increased. Finally, the examination of imagination time length showed that this length could be used as a classification feature. The classification was significantly higher than the level of chance.

Chapter 6 presented the development of a novel feature extraction framework based on examining and reducing the temporal variations between EEG trials using DTW. The classification accuracy of the five imagined words and between speech versus non-speech tasks using the features extracted from the framework were compared. They were examined against time domain (maximum cross-correlation and EEG statistics) and time frequency (relative wavelet energy calculated from discrete wavelet transform coefficients, wavelet transform coefficients and common spatial patterns). The classification accuracy using the proposed framework outperformed the compared features in the classification of speech versus silence. Further, it outperformed time domain features in classifying the five imagined words. Several modifications to the framework were proposed and compared to the main developed framework. These modifications included examining DTW in the same approach used for audio speech

recognition, applying derivative dynamic time warping and applying DTW outlier trials removing. In most cases, the developed DTW framework outperformed the proposed modifications.

7.2 Original contributions and findings

The main scientific contributions resulting from the research reported in this thesis are as follows:

- Recording two EEG datasets for imagined speech. The first dataset was collected from nine subjects. It included imagining 11 randomly presented semantically varying words. In the second dataset, EEG data from 10 subjects were recorded in block mode while they imagined five words.
- Successful discrimination between imagined speech vs non-speech tasks.
  - The results exceeded the chance level for all subjects for the two non-speech tasks: silence time and attention to visual stimuli time.
  - The time of attention to visual stimuli could be classified better than silence time compared to imagined speech, even if this attention was related to two different visual stimuli. The maximum average classification accuracy was 67.15%
  - The results showed the importance of classifying groups of words against non-speech tasks to identify the best type of features for each subject.
  - The current experimental parameters did not show the importance of semantics in improving the classification accuracy. The reasons are discussed in Section 4.3.4.
- Evidence of the importance of optimising the experimental parameters in enhancing the recognition of imagined words was provided. The examined parameters were: an EEG patterns separation method, training size with respect to recording time in the session and the length of the examined time frame.

- Experimental evidence that the imagination time length could be used as a classification feature was given. It yielded 10.55–16.95% above the chance level classification accuracy in classifying five imagined words. This indicated the importance of temporal variation in discriminating imagined speech that is similar to audio-speech.

- A novel feature extraction framework, including using DTW as the first use of imagined speech recognition using EEG was developed and evaluated. The proposed framework significantly outperformed a set of state-of-the-art features in classifying speech versus non-speech and time domain features classifying five imagined words. Moreover, this framework could be generalised for feature extraction for EEG data in any BCI application.

7.3 Future work

7.3.1 Experiments improvement

In discriminating between speech versus non-speech, it was assumed that the semantics of words could help in the recognition of the imagined. However, there was no clear effect of a word's semantics in classifying the imagined words versus non-speech. This could be improved by evoking the emotions using extra visual/audio-visual stimuli, as in EEG, for emotion recognition studies (Murugappan et al., 2010) to generate the emotions connected to the word's meaning.

In the proposed DTW framework, symmetric DTW was applied. It would be interesting to examine the asymmetric DTW in EEG signals. Further work could also be performed on varying the slope condition of the warping path (as described in (Sakoe and Chiba, 1978)).

The proposed DTW feature extraction framework successfully outperformed all the baseline configurations in classifying speech versus silence. However, for the imagined words, the classification statistics of the DWT coefficients significantly outperformed

the proposed DTW framework using the random forest classifier. This would suggest the importance of frequency information to improve classifying imagined words. This could be improved by implementing frequency-based warping. Here, EEG signals would be filtered into several frequency bands before conducting the alignment. However, one crucial step before doing this is understanding the important frequency bands in EEG signals during speech imagination.

7.3.2 Application of the findings

The research in this thesis successfully examined the discrimination of speech versus non-speech and the classification of imagined words. The next step would be proposing a self-paced BCI. Here, the intention of unspoken speech from EEG signals could be detected without requiring extra activities such as mouse clicks or eye blinks. This was partially examined in (Bashashati and Ward, 2017; Song and Sepulveda, 2017) for different speech modes for EEG signals. However, these studies did not involve imagined words classification.

Future work should apply the proposed feature extraction framework to another dataset for other BCI applications such as motor imagination.

Bibliography

Abdallah, N., Daya, B., Khawandi, S., and Chauvet, P. (2017). Electroencephalographic based brain computer interface for unspoken speech. In *Sensors Networks Smart and Emerging Technologies (SENSET)*, pages 1–4. IEEE.

Achanccaray, D., Acuña, K., Carranza, E., and Andreu-Perez, J. (2017). A virtual reality and brain computer interface system for upper limb rehabilitation of post stroke patients. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–5. IEEE.

Alderson-Day, B. and Fernyhough, C. (2015). Inner speech: development, cognitive functions, phenomenology, and neurobiology. *Psychological bulletin*, 141(5):931.

Allison, B., Graimann, B., and Gräser, A. (2007). Why use a bci if you are healthy. In *ACE Workshop-Brain-Computer Interfaces and Games*, pages 7–11.

AlSaleh, M., Moore, R., Christensen, H., and Arvaneh, M. (2018). Discriminating between imagined speech and non-speech tasks using eeg. In *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1952–1955. IEEE.

AlSaleh, M. M., Arvaneh, M., Christensen, H., and Moore, R. K. (2016). Brain-computer interface technology for speech recognition: A review. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA) Asia-Pacific*, pages 1–5. IEEE.

Amiri, S., Rabbi, A., Azinfar, L., and Fazel-Rezai, R. (2013). A review of p300, ssvep, and hybrid p300/ssvep brain-computer interface systems. *Brain-Computer Interface Systems—Recent Progress and Future Prospects*.

Anderson, N. H. (1968). Likableness ratings of 555 personality-trait words. *Journal of personality and social psychology*, 9(3):272.

Ang, K. K., Chin, Z. Y., Wang, C., Guan, C., and Zhang, H. (2012). Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b. *Frontiers in neuroscience*, 6:39.

Ang, K. K., Chin, Z. Y., Zhang, H., and Guan, C. (2008). Filter bank common spatial pattern (fbcsp) in brain-computer interface. In *IEEE International Joint Conference on Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence)*, pages 2390–2397. IEEE.

Angrisani, L., Daponte, P., D'apuzzo, M., and Testa, A. (1998). A measurement method based on the wavelet transform for power quality analysis. *IEEE Transactions on Power Delivery*, 13(4):990–998.

Antonio, T.-G. A., Alberto, R.-G. C., and Luis, V.-P. (2012). Toward a silent speech interface based on unspoken speech. In *Proceeding of the 5th International Joint Conference on Biomedical Engineering Systems and Technologies*.

Bachorowski, J.-A. (1999). Vocal expression and perception of emotion. *Current directions in psychological science*, 8(2):53–57.

Baddeley, A. D., Thomson, N., and Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of verbal learning and verbal behavior*, 14(6):575–589.

Balaji, A., Haldar, A., Patil, K., Ruthvik, T. S., Valliappan, C., Jartarkar, M., and Baths, V. (2017). Eeg-based classification of bilingual unspoken speech using ann. In *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1022–1025. IEEE.

Banse, R. and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614.

Bashashati, H. and Ward, R. (2017). Ensemble of neural network conditional random fields for self-paced brain computer interfaces. *Advances in Science, Technology and Engineering Systems Journal*, 2(3):996–1005.

Basho, S., Palmer, E. D., Rubio, M. A., Wulfeck, B., and Müller, R.-A. (2007). Effects of generation mode in fmri adaptations of semantic fluency: paced production and overt speech. *Neuropsychologia*, 45(8):1697–1706.

Belliveau, J., Kennedy, D., McKinstry, R., Buchbinder, B., Weisskoff, R., Cohen, M., Vevea, J., Brady, T., and Rosen, B. (1991). Functional mapping of the human visual cortex by magnetic resonance imaging. *Science*, 254(5032):716–719.

Bellman, R. (2013). *Dynamic programming*. Courier Corporation.

Bertelson, P., Vroomen, J., and De Gelder, B. (2003). Visual recalibration of auditory speech identification: a mcgurk aftereffect. *Psychological Science*, 14(6):592–597.

Bhavsar, R., Sun, Y., Helian, N., Davey, N., Mayor, D., and Steffert, T. (2018). The correlation between eeg signals as measured in different positions on scalp varying with distance. *Procedia Computer Science*.

Birbaumer, N., Ghanayim, N., Hinterberger, T., Iversen, I., Kotchoubey, B., Kübler, A., Perelmouter, J., Taub, E., and Flor, H. (1999). A spelling device for the paralysed. *Nature*, 398(6725):297.

Birbaumer, N., Hinterberger, T., Kubler, A., and Neumann, N. (2003). The thought-translation device (ttd): neurobehavioral mechanisms and clinical outcome. *IEEE transactions on Neural Systems and rehabilitation engineering*, 11(2):120–123.

Blakely, T., Miller, K. J., Rao, R. P., Holmes, M. D., and Ojemann, J. G. (2008). Localization and classification of phonemes using high spatial resolution electrocorticography (ecog) grids. In *30th Annual International Conference of the Engineering in Medicine and Biology Society (EMBS)*, pages 4964–4967. IEEE.

Blankertz, B., Tangermann, M., Vidaurre, C., Fazli, S., Sannelli, C., Haufe, S., Maeder, C., Ramsey, L. E., Sturm, I., Curio, G., et al. (2010). The berlin brain–computer interface: non-medical uses of bci technology. *Frontiers in neuroscience*, 4:198.

Brigham, K. and Kumar, B. V. (2010a). Imagined speech classification with eeg signals for silent communication: a preliminary investigation into synthetic telepathy. In *4th International Conference on Bioinformatics and Biomedical Engineering (iCBBE)*, pages 1–4. IEEE.

Brigham, K. and Kumar, B. V. (2010b). Subject identification from electroencephalogram (eeg) signals during imagined speech. In *Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS)*, pages 1–8. IEEE.

Broca, P. (1861). Remarques sur le siège de la faculté du langage articulé, suivies d'une observation d'aphémie (perte de la parole). *Bulletin et Memoires de la Societe anatomique de Paris*, 6:330–357.

Brocklehurst, P. H. and Corley, M. (2011). Investigating the inner speech of people who stutter: Evidence for (and against) the covert repair hypothesis. *Journal of Communication Disorders*, 44(2):246–260.

Brumberg, J. S., Wright, E. J., Andreasen, D. S., Guenther, F. H., and Kennedy, P. R. (2011). Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech motor cortex. *Frontiers in neuroscience*, 5:65.

Brunner, C., Leeb, R., Müller-Putz, G., Schlögl, A., and Pfurtscheller, G. (2008). Bci competition 2008–graz data set a. *Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology*, pages 136–142.

Chaovalitwongse, W. and Pardalos, P. (2008). On the time series support vector machine using dynamic time warping kernel for brain activity classification. *Cybernetics and Systems Analysis*, 44(1):125–138.

Chen, X., Wang, Y., Nakanishi, M., Gao, X., Jung, T.-P., and Gao, S. (2015). High-speed spelling with a noninvasive brain–computer interface. *Proceedings of the national academy of sciences*, 112(44):E6058–E6067.

Chia, X., Hagedorna, J. B., Schoonovera, D., and D'Zmuraa, M. (2011). Eeg-based discrimination of imagined speech phonemes. *International Journal of Bioelectromagnetism*, 13(4):201–206.

Combrisson, E. and Jerbi, K. (2015). Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of neuroscience methods*, 250:126–136.

Corley, M., Brocklehurst, P. H., and Moat, H. S. (2011). Error biases in inner and overt speech: Evidence from tongue twisters. *Journal of experimental psychology: Learning, memory, and cognition*, 37(1):162.

Crone, N. E., Boatman, D., Gordon, B., and Hao, L. (2001). Induced electrocorticographic gamma activity during auditory perception. *Clinical neurophysiology*, 112(4):565–582.

Curran, E. A. and Stokes, M. J. (2003). Learning to control brain activity: a review of the production and control of eeg components for driving brain–computer interface (bci) systems. *Brain and cognition*, 51(3):326–336.

Dal Seno, B., Matteucci, M., and Mainardi, L. (2010). Online detection of p300 and error potentials in a bci speller. *Computational intelligence and neuroscience*, 2010:11.

Dalbis, T., Blatt, R., Tedesco, R., Sbattella, L., and Matteucci, M. (2012). A predictive speller controlled by a brain-computer interface based on motor imagery. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 19(3):20.

Daly, J. J., Fang, Y., Perepezko, E. M., Siemionow, V., and Yue, G. H. (2006). Prolonged cognitive planning time, elevated cognitive effort, and relationship to coordination and motor control following stroke. *IEEE Transactions on neural systems and rehabilitation engineering*, 14(2):168–171.

DaSalla, C. S., Kambara, H., Sato, M., and Koike, Y. (2009). Single-trial classification of vowel speech imagery using common spatial patterns. *Neural networks*, 22(9):1334–1339.

Dehaene, S., Posner, M. I., and Tucker, D. M. (1994). Localization of a neural system for error detection and compensation. *Psychological Science*, 5(5):303–305.

Diwaker, S., Gupta, S., and Gupta, N. (2016). Classification of eeg signal using correlation coefficient among channels as features extraction method. *Indian Journal of Science and Technology*, 9(32).

Doerksen, S. and Shimamura, A. P. (2001). Source memory enhancement for emotional words. *Emotion*, 1(1):5.

Dolcos, S. and Albarracin, D. (2014). The inner speech of behavioral regulation: Intentions and task performance strengthen when you talk to yourself as a you. *European Journal of Social Psychology*, 44(6):636–642.

DZmura, M., Deng, S., Lappas, T., Thorpe, S., and Srinivasan, R. (2009). Toward eeg sensing of imagined speech. In *International Conference on Human-Computer Interaction*, pages 40–48. Springer.

Fan, J., Wade, J. W., Key, A. P., Warren, Z. E., and Sarkar, N. (2018). Eeg-based affect and workload recognition in a virtual driving environment for asd intervention. *IEEE Transactions on Biomedical Engineering*, 65(1):43–51.

Fazel-Rezai, R., Allison, B. Z., Guger, C., Sellers, E. W., Kleih, S. C., and Kübler, A. (2012). P300 brain computer interface: current challenges and emerging trends. *Frontiers in neuroengineering*, 5:14.

Ferreira, A. L. S., Marciano, J. N., Miranda, L., and Miranda, E. (2014). Understanding and proposing a design rationale of digital games based on brain-computer interface: Results of the admiralmind battleship study. *SBC Journal on Interactive Systems*, 4(1):3–15.

Fossati, P., Hevenor, S. J., Graham, S. J., Grady, C., Keightley, M. L., Craik, F., and Mayberg, H. (2003). In search of the emotional self: an fmri study using positive and negative emotional words. *American Journal of Psychiatry*, 160(11):1938–1945.

Freeman, W. and Quiroga, R. Q. (2012). *Imaging brain nction with EEG: advanced temporal and spatial analysis of electroencephalographic signals*. Springer Science & Business Media.

Gallagher, A., Thériault, M., Maclin, E., Low, K., Gratton, G., Fabiani, M., Gagnon, L., Valois, K., Rouleau, I., Sauerwein, H. C., et al. (2007). Near-infrared spectroscopy as an alternative to the wada test for language mapping in children, adults and special populations. *Epileptic Disorders*, 9(3):241–255.

Geschwind, N. (1979). Specializations of the human brain. *Scientific American*, 241(3):180–201.

Gick, B., Wilson, I., and Derrick, D. (2012). *Articulatory phonetics*. John Wiley & Sons.

González-Castañeda, E. F., Torres-García, A. A., Reyes-García, C. A., and Villaseñor-Pineda, L. (2017). Sonification and textification: Proposing methods for classifying unspoken words from eeg signals. *Biomedical Signal Processing and Control*, 37:82–91.

Guenther, F. H. and Brumberg, J. S. (2011). Brain-machine interfaces for real-time speech synthesis. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual international conference of the IEEE*, pages 5360–5363. IEEE.

Gui, Q., Jin, Z., Blondet, M. V. R., Laszlo, S., and Xu, W. (2015). Towards eeg biometrics: similarity-based approaches for user identification. In *IEEE International Conference on Identity, Security and Behavior Analysis*, pages 1–6.

Guimaraes, M. P., Wong, D. K., Uy, E. T., Grosenick, L., and Suppes, P. (2007). Single-trial classification of meg recordings. *IEEE Transactions on Biomedical Engineering*, 54(3):436–443.

Guo, L., Rivero, D., Seoane, J. A., and Pazos, A. (2009). Classification of eeg signals using relative wavelet energy and artificial neural networks. In *Proceedings of the first ACM/SIGEVO Summit on Genetic and Evolutionary Computation*, pages 177–184. ACM.

Gupta, S. S., Soman, S., Raj, P. G., Prakash, R., Sailaja, S., and Borgohain, R. (2012). Detecting eye movements in eeg for controlling devices. In *IEEE International Conference on Computational Intelligence and Cybernetics (CyberneticsCom)*, pages 69–73. IEEE.

Heinks-Maldonado, T. H., Nagarajan, S. S., and Houde, J. F. (2006). Magnetoencephalographic evidence for a precise forward model in speech production. *NeuroReport*, 17(13):1375.

Herbert, C., Junghofer, M., and Kissler, J. (2008). Event related potentials to emotional adjectives during reading. *Psychophysiology*, 45(3):487–498.

Herff, C., Heger, D., Putze, F., Guan, C., and Schultz, T. (2013). Self-paced bci with nirs based on speech activity. In *International BCI Meeting*.

Herff, C., Putze, F., Heger, D., Guan, C., and Schultz, T. (2012). Speaking mode recognition from functional near infrared spectroscopy. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1715–1718. IEEE.

Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. *Trends in cognitive sciences*, 6(6):242–247.

Hong, B., Guo, F., Liu, T., Gao, X., and Gao, S. (2009). N200-speller using motion-onset visual response. *Clinical neurophysiology*, 120(9):1658–1666.

Houde, J. F., Nagarajan, S. S., Sekihara, K., and Merzenich, M. M. (2002). Modulation of the auditory cortex during speech: an meg study. *Journal of cognitive neuroscience*, 14(8):1125–1138.

Huang, R., Liu, Q., Lu, H., and Ma, S. (2002). Solving the small sample size problem of lda. In *Proceeding of the 16th International Conference on Pattern Recognition.*, volume 3, pages 29–32. IEEE.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453.

Hwang, H.-J., Choi, H., Kim, J.-Y., Chang, W.-D., Kim, D.-W., Kim, K., Jo, S., and Im, C.-H. (2016). Toward more intuitive brain–computer interfacing: classification of binary covert intentions using functional near-infrared spectroscopy. *Journal of biomedical optics*, 21(9):091303.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430.

Idrees, B. M. and Farooq, O. (2016). Vowel classification using wavelet decomposition during speech imagery. In *3rd International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 636–640. IEEE.

Indefrey, P. and Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition*, 92(1-2):101–144.

Jahangiri, A., Chau, J. M., and Achanccaray, D. R. (2018). Covert Speech vs . Motor Imagery : a comparative study of class separability in identical environments. In *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2020–2023. IEEE.

Jahangiri, A. and Sepulveda, F. (2017). The contribution of different frequency bands in class separability of covert speech tasks for bcis. In *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2093–2096. IEEE.

Jasper, H. H. (1958a). The ten-twenty electrode system of the international federation. *Electroencephalogy and Clinical Neurophysiology*, 10:370–375.

Jasper, H. H. (1958b). The ten twenty electrode system of the international federation. *Electroencephalography and clinical neurophysiology*, 10:371–375.

Jauberteau, F. and Jauberteau, J. (2009). Numerical differentiation with noisy signal. *Applied mathematics and computation*, 215(6):2283–2297.

Jobsis, F. F. (1977). Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. *Science*, 198(4323):1264–1267.

John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proceedings of the 11th conference on Uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc.

Karamzadeh, N., Medvedev, A., Azari, A., Gandjbakhche, A., and Najafizadeh, L. (2013). Capturing dynamic patterns of task-based functional connectivity with eeg. *Neuroimage*, 66:311–317.

Kawanabe, M., Pascual, J., and Vidaurre, C. (2010). Investigation of non-stationarity in brain activity via robust principal component analysis. *Frontiers in Computational Neuroscience*.

Kellis, S., Miller, K., Thomson, K., Brown, R., House, P., and Greger, B. (2010). Decoding spoken words using local field potentials recorded from the cortical surface. *Journal of neural engineering*, 7(5):056007.

Keogh, E. J. and Pazzani, M. J. (2001). Derivative dynamic time warping. In *Proceedings of the 2001 SIAM International Conference on Data Mining*, pages 1–11. SIAM.

Komandur, S., Johnson, P. W., and Storch, R. (2008). Relation between mouse button click duration and muscle contraction time. In *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, pages 2299–2301. IEEE.

Krishna, D. H., Pasha, I., and Savithri, T. S. (2016). Classification of eeg motor imagery multi class signals based on cross correlation. *Procedia Computer Science*, 85:490–495.

Kryger, M., Wester, B., Pohlmeyer, E. A., Rich, M., John, B., Beaty, J., McLoughlin, M., Boninger, M., and Tyler-Kabara, E. C. (2017). Flight simulation using a brain-computer interface: A pilot, pilot study. *Experimental neurology*, 287:473–478.

Kubler, A., Mushahwar, V., Hochberg, L. R., and Donoghue, J. P. (2006). Bci meeting 2005-workshop on clinical issues and applications. *IEEE Transactions on neural systems and rehabilitation engineering*, 14(2):131–134.

Kumagai, Y., Arvaneh, M., Okawa, H., Wada, T., and Tanaka, T. (2017). Classification of familiarity based on cross-correlation features between eeg and music. In *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2879–2882. IEEE.

Kumar, P., Saini, R., Roy, P. P., Sahu, P. K., and Dogra, D. P. (2018). Envisioned speech recognition using eeg sensors. *Personal and Ubiquitous Computing*, 22(1):185–199.

Leuthardt, E., Pei, X.-M., Breshears, J., Gaona, C., Sharma, M., Freudenburg, Z., Barbour, D., and Schalk, G. (2012). Temporal evolution of gamma activity in human cortex during an overt and covert word repetition task. *Frontiers in human neuroscience*, 6:99.

Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., and Yger, F. (2018). A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update. *Journal of neural engineering*, 15(3):031005.

Lotte, F., Brumberg, J. S., Brunner, P., Gunduz, A., Ritaccio, A. L., Guan, C., and Schalk, G. (2015). Electrocorticographic representations of segmental features in continuous speech. *Frontiers in human neuroscience*, 9.

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., and Arnaldi, B. (2007). A review of classification algorithms for eeg-based brain–computer interfaces. *Journal of neural engineering*, 4(2):R1.

Lotte, F., Faller, J., Guger, C., Renard, Y., Pfurtscheller, G., Lécuyer, A., and Leeb, R. (2012). Combining bci with virtual reality: towards new applications and improved bci. In *Towards Practical Brain-Computer Interfaces*, pages 197–220. Springer.

MacKay, D. G. (2014). Constraints on theories of inner speech. In *Auditory imagery*, pages 133–162. Psychology Press.

Mak, J. N. and Wolpaw, J. R. (2009). Clinical applications of brain-computer interfaces: current state and future prospects. *IEEE reviews in biomedical engineering*, 2:187.

Markopoulos, P. P. (2017). Linear discriminant analysis with few training data. In *Proceeding of the Acoustics, Speech and Signal Processing (ICASSP)*, pages 4626–4630. IEEE.

Martin, S., Brunner, P., Iturrate, I., Millán, J. d. R., Schalk, G., Knight, R. T., and Pasley, B. N. (2016). Word pair classification during imagined speech using direct brain recordings. *Scientific reports*, 6:25803.

Martínez-Manrique, F. and Vicente, A. (2015). The activity view of inner speech. *Frontiers in psychology*, 6:232.

Marvel, C. L. and Desmond, J. E. (2012). From storage to manipulation: how the neural correlates of verbal working memory reflect varying demands on inner speech. *Brain and language*, 120(1):42–51.

Matsumoto, M. and Hori, J. (2013). Classification of silent speech using adaptive collection. In *IEEE Symposium on Computational Intelligence in Rehabilitation and Assistive Technologies (CIRAT)*, pages 5–12. IEEE.

Matsumoto, M. and Hori, J. (2014). Classification of silent speech using support vector machine and relevance vector machine. *Applied Soft Computing*, 20:95–102.

McFarland, D. J., McCane, L. M., David, S. V., and Wolpaw, J. R. (1997). Spatial filter selection for eeg-based communication. *Electroencephalography and clinical Neurophysiology*, 103(3):386–394.

McGuire, P., Silbersweig, D., Murray, R., David, A., Frackowiak, R., and Frith, C. (1996). Functional anatomy of inner speech and auditory verbal imagery. *Psychological medicine*, 26(1):29–38.

Millán, J. d. R., Rupp, R., Müller-Putz, G., Murray-Smith, R., Giugliemma, C., Tangermann, M., Vidaurre, C., Cincotti, F., Kubler, A., Leeb, R., et al. (2010). Combining brain–computer interfaces and assistive technologies: state-of-the-art and challenges. *Frontiers in neuroscience*, 4:161.

Miller, N., Maruyama, G., Beaber, R. J., and Valone, K. (1976). Speed of speech and persuasion. *Journal of personality and social psychology*, 34(4):615.

Min, B., Kim, J., Park, H.-j., and Lee, B. (2016). Vowel imagery decoding toward silent speech bci using extreme learning machine with electroencephalogram. *BioMed research international*, 2016.

Min, B.-K., Marzelli, M. J., and Yoo, S.-S. (2010). Neuroimaging-based approaches in the brain–computer interface. *Trends in biotechnology*, 28(11):552–560.

Misaki, M., Kim, Y., Bandettini, P. A., and Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fmri. *Neuroimage*, 53(1):103–118.

Morin, A. and Michaud, J. (2007). Self-awareness and the left inferior frontal gyrus: inner speech use during self-related processing. *Brain research bulletin*, 74(6):387–396.

Mugler, E. M., Patton, J. L., Flint, R. D., Wright, Z. A., Schuele, S. U., Rosenow, J., Shih, J. J., Krusienski, D. J., and Slutzky, M. W. (2014). Direct classification of all american english phonemes using signals from functional speech motor cortex. *Journal of neural engineering*, 11(3):035015.

Müller, K.-R., Krauledat, M., Dornhege, G., Curio, G., and Blankertz, B. (2004). Machine learning techniques for brain-computer interfaces.

Murugappan, M., Ramachandran, N., and Sazali, Y. (2010). Classification of human emotion from eeg using discrete wavelet transform. *Journal of Biomedical Science and Engineering*, 3(04):390.

Naci, L., Cusack, R., Jia, V. Z., and Owen, A. M. (2013). The brain's silent messenger: using selective attention to decode human thought for brain-based communication. *Journal of Neuroscience*, 33(22):9385–9393.

Nguyen, C. H., Karavas, G. K., and Artemiadis, P. (2017). Inferring imagined speech using eeg signals: a new approach using riemannian manifold features. *Journal of neural engineering*, 15(1):016002.

Nunez, P. L., Srinivasan, R., Westdorp, A. F., Wijesinghe, R. S., Tucker, D. M., Silberstein, R. B., and Cadusch, P. J. (1997). Eeg coherency: I: statistics, reference electrode, volume conduction, laplacians, cortical imaging, and interpretation at multiple scales. *Electroencephalography and clinical neurophysiology*, 103(5):499–515.

Oken, B. S., Orhan, U., Roark, B., Erdogmus, D., Fowler, A., Mooney, A., Peters, B., Miller, M., and Fried-Oken, M. B. (2014). Brain–computer interface with language model–electroencephalography fusion for locked-in syndrome. *Neurorehabilitation and neural repair*, 28(4):387–394.

Oppenheim, G. M. and Dell, G. S. (2008). Inner speech slips exhibit lexical bias, but not the phonemic similarity effect. *Cognition*, 106(1):528–537.

Oppenheim, G. M. and Dell, G. S. (2010). Motor movement matters: The flexible abstractness of inner speech. *Memory & cognition*, 38(8):1147–1160.

Ou, C.-Z., Lin, B.-S., Chang, C.-J., and Lin, C.-T. (2012). Brain computer interface-based smart environmental control system. In *Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*, pages 281–284. IEEE.

Palmer, E. D., Rosen, H. J., Ojemann, J. G., Buckner, R. L., Kelley, W. M., and Petersen, S. E. (2001). An event-related fmri study of overt and covert word stem completion. *Neuroimage*, 14(1):182–193.

Pardey, J., Roberts, S., and Tarassenko, L. (1996). A review of parametric modelling techniques for eeg analysis. *Medical engineering & physics*, 18(1):2–11.

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., Knight, R. T., and Chang, E. F. (2012). Reconstructing speech from human auditory cortex. *PLoS biology*, 10(1):e1001251.

Paulesu, E., Frith, C. D., and Frackowiak, R. S. (1993). The neural correlates of the verbal component of working memory. *Nature*, 362(6418):342.

Pei, X., Hill, J., and Schalk, G. (2012). Silent communication: toward using brain signals. *IEEE pulse*, 3(1):43–46.

Perdikis, S., Leeb, R., Williamson, J., Ramsay, A., Tavella, M., Desideri, L., Hoogerwerf, E.-J., Al-Khodairy, A., Murray-Smith, R., and d R Millán, J. (2014). Clinical evaluation of braintree, a motor imagery hybrid bci speller. *Journal of neural engineering*, 11(3):036003.

Perrone-Bertolotti, M., Rapin, L., Lachaux, J.-P., Baciu, M., and Loevenbruck, H. (2014). What is that little voice inside my head? inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring. *Behavioural brain research*, 261:220–239.

Pfurtscheller, G. and Da Silva, F. L. (1999). Event-related eeg/meg synchronization and desynchronization: basic principles. *Clinical neurophysiology*, 110(11):1842–1857.

Porbadnigk, A., Wester, M., and Jan-p Calliess, T. S. (2009). Eeg-based speech recognition impact of temporal effects.

Qureshi, M. N. I., Min, B., Park, H.-j., Cho, D., Choi, W., and Lee, B. (2018). Multiclass classification of word imagination speech with hybrid connectivity features. *IEEE Transactions on Biomedical Engineering*, 65(10):2168–2177.

Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., Zakaria, J., and Keogh, E. (2012). Searching and mining trillions of time series subsequences under dynamic time warping. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 262–270. ACM.

Ramoser, H., Muller-Gerking, J., and Pfurtscheller, G. (2000). Optimal spatial filtering of single trial eeg during imagined hand movement. *IEEE transactions on rehabilitation engineering*, 8(4):441–446.

Rampp, S. and Stefan, H. (2007). On the opposition of eeg and meg. *Clinical neurophysiology*, 118(8):1658–1659.

Rao, R. P. and Scherer, R. (2010). Brain-computer interfacing [in the spotlight]. *IEEE Signal Processing Magazine*, 27(4):152–150.

Rebsamen, B., Guan, C., Zhang, H., Wang, C., Teo, C., Ang, M. H., and Burdet, E. (2010). A brain controlled wheelchair to navigate in familiar environments. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 18(6):590–598.

Rezazadeh Sereshkeh, A., Trott, R., Bricout, A., and Chau, T. (2017). Eeg classification of covert speech using regularized neural networks. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 25(12):2292–2300.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49.

Sarmiento, L., Lorenzana, P., Cortes, C., Arcos, W., Bacca, J., and Tovar, A. (2014). Brain computer interface (bci) with eeg signals for automatic vowel recognition based on articulation mode. In *5th ISSNIP-IEEE Biosignals and Biorobotics Conference (2014): Biosignals and Robotics for Better and Safer Living (BRC)*, pages 1–4. IEEE.

Schacht, A. and Sommer, W. (2009). Time course and task dependence of emotion effects in word processing. *Cognitive, Affective, & Behavioral Neuroscience*, 9(1):28–43.

Schepers, I. M., Yoshor, D., and Beauchamp, M. S. (2014). Electrocorticography reveals enhanced visual cortex responses to visual speech. *Cerebral Cortex*, 25(11):4103–4110.

Sereshkeh, A. R., Trott, R., Bricout, A., and Chau, T. (2017a). Eeg classification of covert speech using regularized neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2292–2300.

Sereshkeh, A. R., Trott, R., Bricout, A., and Chau, T. (2017b). Online eeg classification of covert speech for brain–computer interfacing. *International journal of neural systems*, 27(08):1750033.

Shenoy, P., Krauledat, M., Blankertz, B., Rao, R. P., and Müller, K.-R. (2006). Towards adaptive classification for bci. *Journal of neural engineering*, 3(1):R13.

Song, Y. and Sepulveda, F. (2017). A novel onset detection technique for brain–computer interfaces using sound-production related cognitive tasks in simulated-online system. *Journal of neural engineering*, 14(1):016019.

Stark, B. C., Geva, S., and Warburton, E. A. (2017). Inner speech's relationship with overt speech in poststroke aphasia. *Journal of Speech, Language, and Hearing Research*, 60(9):2406–2415.

Subasi, A. (2007). Eeg signal classification using wavelet feature extraction and a mixture of expert model. *Expert Systems with Applications*, 32(4):1084–1093.

Suppes, P., Lu, Z.-L., and Han, B. (1997). Brain wave recognition of words. *Proceedings of the National Academy of Sciences*, 94(26):14965–14969.

Sutton, S., Braren, M., Zubin, J., and John, E. (1965). Evoked-potential correlates of stimulus uncertainty. *Science*, 150(3700):1187–1188.

Teplan, M. et al. (2002). Fundamentals of eeg measurement. *Measurement science review*, 2(2):1–11.

Tervaniemi, M. a., Kujala, A., Alho, K., Virtanen, J., Ilmoniemi, R., and Näätänen, R. (1999). Functional specialization of the human auditory cortex in processing phonetic and musical sounds: a magnetoencephalographic (meg) study. *Neuroimage*, 9(3):330–336.

Torres-García, A. A., Reyes-García, C. A., Villaseñor-Pineda, L., and García-Aguilar, G. (2016). Implementing a fuzzy inference system in a multi-objective eeg channel selection model for imagined speech classification. *Expert Systems with Applications*, 59:1–12.

Trambaiolli, L. R. and Falk, T. H. (2018). Hybrid brain–computer interfaces for wheelchair control: a review of existing solutions, their advantages and open challenges. In *Smart Wheelchairs and Brain-Computer Interfaces*, pages 229–256. Elsevier.

Vaadia, E. and Birbaumer, N. (2009). Grand challenges of brain computer interfaces in the years to come. *Frontiers in neuroscience*, 3:15.

Van Gerven, M., Farquhar, J., Schaefer, R., Vlek, R., Geuze, J., Nijholt, A., Ramsey, N., Haselager, P., Vuurpijl, L., Gielen, S., et al. (2009). The brain–computer interface cycle. *Journal of neural engineering*, 6(4):041001.

Vanacker, G., Millán, J. d. R., Lew, E., Ferrez, P. W., Moles, F. G., Philips, J., Van Brussel, H., and Nuttin, M. (2007). Context-based filtering for assisted brain-actuated wheelchair driving. *Computational intelligence and neuroscience*, 2007.

Wang, J., Kim, M., Hernandez-Mulero, A. W., Heitzman, D., and Ferrari, P. (2017). Towards decoding speech production from single-trial magnetoencephalography (meg) signals. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3036–3040. IEEE.

Wang, M., Daly, I., Allison, B. Z., Jin, J., Zhang, Y., Chen, L., and Wang, X. (2015). A new hybrid bci paradigm based on p300 and ssvep. *Journal of neuroscience methods*, 244:16–25.

Watanabe, E., Maki, A., Kawaguchi, F., Takashiro, K., Yamashita, Y., Koizumi, H., and Mayanagi, Y. (1998). Non-invasive assessment of language dominance with near-infrared spectroscopic mapping. *Neuroscience letters*, 256(1):49–52.

Wernicke, C. (1974). Der aphasische symptomenkomplex. In *Der aphasische Symptomencomplex*, pages 1–70. Springer.

Wester, M. (2006). Unspoken speech-speech recognition based on electroencephalography. *Master's Thesis, Universitat Karlsruhe (TH)*.

Wheeldon, L. R. and Levelt, W. J. (1995). Monitoring the time course of phonological encoding.

Wolpaw, J. R., Birbaumer, N., Heetderks, W. J., McFarland, D. J., Peckham, P. H., Schalk, G., Donchin, E., Quatrano, L. A., Robinson, C. J., and Vaughan, T. M. (2000). Brain-computer interface technology: a review of the first international meeting. *IEEE transactions on rehabilitation engineering*, 8(2):164–173.

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., and Vaughan, T. M. (2002). Brain–computer interfaces for communication and control. *Clinical neurophysiology*, 113(6):767–791.

Woodman, G. F. (2010). A brief introduction to the use of event-related potentials in studies of perception and attention. *Attention, Perception, & Psychophysics*, 72(8):2031–2046.

Yang, Z., Huang, Z., Gonzalez-Castillo, J., Dai, R., Northoff, G., and Bandettini, P. (2014). Using fmri to decode true thoughts independent of intention to conceal. *NeuroImage*, 99:80–92.

Yohanes, R. E., Ser, W., and Huang, G.-b. (2012). Discrete wavelet transform coefficients for emotion recognition from eeg signals. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2251–2254. IEEE.

Yoshimura, N., Satsuma, A., DaSalla, C. S., Hanakawa, T., Sato, M.-a., and Koike, Y. (2011). Usability of eeg cortical currents in classification of vowel speech imagery. In *International Conference on Virtual Rehabilitation (ICVR)*, pages 1–2. IEEE.

Zhang, D., Gong, E., Wu, W., Lin, J., Zhou, W., and Hong, B. (2012). Spoken sentences decoding based on intracranial high gamma response using dynamic time warping. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3292–3295. IEEE.

Zhang, R., Wang, Q., Li, K., He, S., Qin, S., Feng, Z., Chen, Y., Song, P., Yang, T., Zhang, Y., et al. (2017). A bci-based environmental control system for patients with severe spinal cord injuries. *IEEE Transactions on Biomedical Engineering*, 64(8):1959–1971.

Zhao, S. and Rudzicz, F. (2015). Classifying phonological categories in imagined and articulated speech. In *Proceeding of the Acoustics, Speech and Signal Processing (ICASSP)*, pages 992–996. IEEE.

Zoumpoulaki, A., Alsufyani, A., Filetti, M., Brammer, M., and Bowman, H. (2015). Latency as a region contrast: Measuring erp latency differences with dynamic time warping. *Psychophysiology*, 52(12):1559–1576.

Appendix A

Data Recording Forms

This appendix lists the data recording forms. The following documents were provided to each participant on the recording day:

- The '**Information Sheet**' shows the recording instructions and explains to the participant how to perform the task.
- The '**Screening Form**' form is used to collect participants personal information. In this form all the questions are designed to ensure the sutibalilty of the participant to perform the experiment.
- Each participant signed and dated two copies of the '**Consent Form**' on the day of recording.

The University Of Sheffield.

Information Sheet

A study about speech imagination using EEG signals

Researchers

Lead Researcher:

Mashael AlSaleh (mmalsaleh1@sheffield.ac.uk)

Supervisors:

Prof. Roger Moore: r.k.moore@sheffield.ac.uk

Dr. Mahnaz Arvaneh: m.arvaneh@sheffield.ac.uk

Invitation

You are being invited to take part in a research study. Before you decide whether you want to take part in this study, you need to understand why we are doing this research and what it will involve. Please read the information below carefully and contact us if there is anything you don't understand or if you would like more information.

Aim

We would like to learn more about the brain's electrical activity when we imagine the pronunciation of words. Previous researches has tried to prove the ability of recognizing and distinguishing brain activities during the actual or imagined speech process. We intend to explore whether a similar recognition can be done with understanding for brain regions that are activated when we intend to speak. The findings may have useful implications toward brain-speech recognition system.

Why have I been chosen?

We are asking healthy young people and adults to take part in this study. If you have a current diagnosis of any neurological or psychiatric condition you are unable to take part in the study. Before taking part in this study, you will need to complete a brief medical and family history to make sure you are able to participate.

Do I have to take part?

It is up to you to decide whether or not to take part. If you do decide to take part you will be given this information sheet to keep, and will be asked to sign a consent form.

What will happen to me if I take part?

We will use a technique called electroencephalography (EEG), which will measure your brain's electrical activity. EEG is a non-invasive and very safe technique with no

direct known health risk. Throughout the computer task you will wear a cap of electrodes, which will record the electrical activity from your brain. The electrodes will be filled with a salt-based gel, which can be easily washed out with wipe/shampoo and does not have any effect on hair color. No allergic reaction known previously from the use of this gel.
The electrode cap will take approximately 15 minutes to set up and the computer task will last around 40 minutes. There will be opportunities to take breaks every 7 minutes. Overall, the study will last not more than one hour and a half. You are free to withdraw from this experiment at any time and do not need to give a reason for doing so.

Time Commitment

The session will be one hour and a half long and the participant is asked to attend once for the experiment.

What are the possible benefits of taking part?

It does not provide any tangible benefits beyond advancing scientific knowledge.

What happens if the research study stops earlier than expected?

You may decide to stop being a part of the research study at any time without explanation. You have the right to ask that any data you have supplied to that point be
withdrawn/destroyed. You have the right to have your questions about the procedures answered (unless answering these questions would interfere with the study's outcome). If you have any questions as a result of reading this information sheet, you should ask the researcher before the study begins.

What are the possible disadvantages and risks of taking part?

No possible disadvantages or risks are envisaged. However, if, at any point during the experiment, you decide that you do not want to carry on, we will stop the experiment and you are free to withdraw from the study without giving a reason.

What if something goes wrong?

In the first instance you should contact the Principal Investigator (contact details are given at the end of this document) should you wish to raise a complaint. However, if you feel your complaint has not been handled to your satisfaction you can contact the Head of Department, who will then escalate the complaint through the appropriate channels.

What will happen to the results of the research project?

The outcome of this study may form part of one or more scientific publications; you will

be entitled to copies of any such publications. You will not be identified in any report or publication. The data collected during the course of this study might be used for additional or subsequent research.

Who is organising and funding the research?

This research is supported financially by the Saudi Ministry of Education.

Who has ethically reviewed the project?

This project has been ethically approved via the University of Sheffield's ethics review procedure.

Confidentiality/Anonymity

Before taking part in this study, you will be asked to provide your name, gender, date of birth and some medical information. Your personal details will be stored in a locked filing cabinet and on a password-protected computer. All information collected from this study is confidential. We will make sure that your information is kept confidential by using identification numbers in place of your name. This will make sure that your name will not be associated with, or traceable to, any of the collected data. The file will be maintained by the lead researcher. The results from this study may be used anonymously at conferences and written up in scientific journals.

Cost and Compensation

You will be entered into a prize draw to win one 20 GBP Amazon voucher.

The University Of Sheffield.

PARTICIPANT SCREENING FORM

CONFIDENTIALITY - This form and the information contained within it will be treated as a confidential document.

Personal Details	
First name	
Last name	
Email address	
Date of birth	
Gender (Male, Female)	
Handedness (Right, Left, Ambidextrous)	
Have you been in an EEG study before? How many times?	

Please answer <u>ALL</u> of the following questions.

Please circle the appropriate answer.

Medical History		
Do you suffer from epilepsy, blackouts, fainting turns or unexplained loss of consciousness, or recurrent headaches?	Yes	No
Do you have family history of epilepsy?	Yes	No
Have you suffered a head injury leading to loss of consciousness requiring a hospital admission?	Yes	No
Do you suffer from any other medical condition, including heart problems?	Yes	No
Do you have a heart or neural pacemaker?	Yes	No
Are you currently taking any prescribed drugs?	Yes	No

Medical History		
Do you currently use (the past 12 hours) any recreational drugs, or have you had problems with alcohol or drug addiction in the past?	Yes	No
What is your visual correction? (please tick) • Uncorrected vision (no glasses or contacts) • Glasses and contacts • Contacts only • Glasses only & can see at arm's length without them • Glasses only & cannot see at arm's length without them		
Do you have any other problems with your sight (e.g. scotoma, colour blindness, blindness in one eye, night blindness, reduced visual field, blurred vision, or detached retina)?	Yes	No
Do you have any problems with your hearing?	Yes	No
Do you wear a hearing aid?	Yes	No
Are there any other medical conditions we should know about?	Yes	No
Details:		

Thank you for completing this screening form.

The University Of Sheffield.

CONSENT FORM

A study about speech imagination using EEG signals

The participants should complete this consent form themselves.

Please read the following statements and circle the appropriate answer.

Have you read the Participant Information Sheet?	Yes	No
Have you had an opportunity to ask questions and discuss the study?	Yes	No
Have you received satisfactory answers to all of your questions?	Yes	No
Have you received enough information about the study?	Yes	No
Do you understand that you are free to withdraw from the study at any time and without having to give a reason for withdrawing?	Yes	No
Who have you spoken to? Dr/Mr/Mrs/Miss		
Do you agree to take part in this study?	Yes	No

Signed: ..... (Participant)

Print name: ..... (Participant)

Date: .....