Decoding Auditory Attention from Multivariate Neural Data using Cepstral Analysis

Carlos Francisco Mendoza & Andrew Segar

Supervisor: Maria Sandsten

Department of Mathematical Statistics
Lund University

This thesis is submitted for the degree of
Master of Science

June 2018

Acknowledgements

We would like to thank Maria Sandsten, our supervisor at Lund University, who has inspired and encouraged us throughout this project. Maria's guidance has been invaluable.

This masters thesis was conducted as a collaboration with Eriksholm Research Centre. We thank our supervisors at Eriksholm; Emina Alickovic and Carina Graversen, who provided so much insight. We were continually boosted by their extensive experience and enthusiasm for their work. Our sincere gratitude goes to Thomas Lunner for pursuing and investing in this collaboration. We are grateful to Thomas involving us in numerous events at Eriksholm, making us feel valued as members of the team, and for expressing interest in us as individuals. We would also like to thank all of the staff at Eriksholm Research Centre for welcoming and including us.

We thank Magnus Wiktorsson for his comments on the final version of the thesis as well as his support and advice throughout the master's programme. We also thank Maria Lövgren, the programme administrator, for her assistance and kindness throughout.

Our gratitude goes to all professors and teaching staff in the Mathematical Statistics department. Their excellent advice and teaching made our studies at Lund University an extremely enriching experience.

Finally we would like to express our gratitude to the Academic Support Centre, particularly Ladaea Rylander for helping us with the writing structure and overall layout of the thesis.

Carlos Francisco Mendoza and Andrew Segar
Lund, Sweden, June 2018

I would like to thank Consejo Nacional de Ciencia y Tecnología (CONACYT) who sponsored this opportunity to study at Lund University. I wish to acknowledge my alma mater; Universidad Nacional Autónoma de México (UNAM) for all the learning that it provided during my years there.

To my parents; María Eugenia Lara and Francisco Mendoza, to my siblings; Lorena, Angélica, and Daniel, and to my beloved life partner; Montserrat Arnaiz, I will be forever grateful for their infinite love and support and for always being my compass.

Carlos Francisco Mendoza
Lund, Sweden, June 2018

I thank my friends and family for all of their help and support.

Andrew Segar
Lund, Sweden, June 2018

Abstract

Very little is known about the remarkable ability of humans to separate a single sound source from a dense mixture of sound sources in a crowded background, known as the cocktail-party scenario. Better understanding could lead to a breakthrough for the next-generation of hearing aids to have the ability to be cognitively controlled. A key finding in the field is that human cortical activity has been shown to follow the speech envelope. However, in these experimental results, the correlation coefficients between the EEG and speech envelope are very low, on the order of r = 0.1-0.2. Also, classification rates are not yet 100%.

The aim of this project is to investigate whether cepstral analysis can be used as a more robust mapping between speech and EEG. Our preliminary results show correlations on the order of r > 0.5. This thesis will give a insight into the method we are developing, our current results, and the expected future results and applications in hearing aids.

Keywords; *signal processing, hearing care, cocktail-party problem, neuroscience, speech processing, spectral analysis, cepstral analysis, cepstrum, stimulus-reconstruction.*

Popular Summary

Very little is known about the remarkable ability of humans to separate a single sound source from a dense mixture of sound sources in a crowded background, known as the cocktail-party scenario. In these situations, those of us fortunate enough to have normal hearing are usually able to tune into a particular speaker with little effort. However if there is background noise, or the listener has a hearing impairment, this can be extremely difficult.

Most people with hearing impairments find it far more difficult to understand speech in noisy environments compared to speech in quiet environments and standard hearing aids are of little help to those with this type of hearing impairment as they amplify both speech and noise. Directional hearing aids provide increases in the signal-to-noise ratio (SNR) which improves comprehension. However, listeners need to face the signal of interest and be within a certain distance to obtain directional benefit.

One of the goals of current research is to create a system that can decide which sound source a listener is attending to, and then steer the directional microphones and/or suppress noise. One method being attempted is to use electroencephalogram (EEG) signals to create a brain-computer interface (BCI) system.

A key finding in the field is that human cortical activity has been shown to follow the speech envelope. Revealing this connection between the envelope of speech and the neural response has opened up the possibility of applying other speech processing techniques to the problem of determining listening attention. Originally developed in the early 1960s as a method of detecting echoes in a signal, cepstral analysis was soon found to be useful for speech processing due to the characteristics of human speech.

The utility of cepstral analysis in speech processing meant it was natural for it to be considered in this problem. There is currently no published research on the application of cepstral analysis to find connections between speech and the neural response. This project was intended as a preliminary investigation into the possibility of using cepstral processing to determine listening attention.

This is a preliminary investigation into the possibilities of using a BCI system to determine listening attention for use in future hearing technology, thus the data used is from an experimental setup with a very abstract scenario. Real-life scenarios are likely to bring

a multitude of other complexities to this problem, however, at this early stage of research simplistic scenarios are essential in order to break the problem into more manageable steps.

Building on the research of O'Sullivan (2014), cepstral coefficients were incorporated into a stimulus-reconstruction model to see if those results could be replicated or even improved using cepstral analysis. The results showed classification rates of over 90% which is an improvement, suggesting that cepstral analysis may be a more effective method than using the speech envelope although further investigation using other data sets is required to confirm this.

In summary, having developed a method of determining listening attention by incorporating cepstral processing techniques into a stimulus-reconstruction model, we have been able to achieve good classification rates. This suggests that cepstral analysis, a technique that has proved useful in speech analysis, can be used to distinguish between attended and unattended speech, potentially adding a new tool for future attempts at determining listening attention in multi-speaker environments. It should be noted however, that further investigation and application of the model to other datasets is required before firm conclusions can be drawn.

Contents

Chapter 1

Introduction

We communicate through speech so often and so effortlessly that it makes it hard to appreciate how complicated the act of listening is. We become more aware of the limitations of our hearing when we are in environments where many people are talking simultaneously [7].

People's ability to separate multiple speech streams was first investigated in the 1950's. In 1953 Colin Cherry published a paper titled "Some experiments on the recognition of speech, in one and two ears", which describes a number of experiments on speech recognition. Included were experiments relating to the separation of two simultaneously spoken messages, looking at the behavior of a listener when presented with two speech signals simultaneously. This was the first attempt to address the question "how do we recognise what one person is saying when others are speaking at the same time?", which Cherry called the "cocktail party problem". Cherry's paper inspired research in a wide range of areas related to selective listening.

In "cocktail party" situations, as demonstrated in Cherry's paper, those of us fortunate enough to have normal hearing are usually able to tune into a particular speaker with little effort. However if there is background noise, or the listener has a hearing impairment, this can be extremely difficult [6]. Most people with hearing impairments find it far more difficult to understand speech in noisy environments compared to speech in quiet environments [6]. Standard hearing aids are of little help to those with this type of hearing impairment as they amplify both speech and noise.

Directional hearing aids provide increases in the signal-to-noise ratio (SNR) which improves comprehension. However, in general directional hearing aids are designed to reduce sounds that are not directly in front of the listener, which creates limitations. Data has shown that listeners need to face the signal of interest and be within a certain distance to obtain directional benefit. Also, noise should either surround, or be directly behind, the listener [25]. One of the goals of current research is to create a system that can decide which sound source a listener is attending to, and then steer the directional microphones and/or suppress noise Wöstmann et al. [32]. One method being attempted is to use electroencephalogram (EEG) signals to create a brain-computer interface (BCI) system.

A key finding is that there is a connection between the neural response and the envelope of the speech stimulus. This finding has been used to show that the M/EEG response shows a stronger correlation with the envelope of attended speech compared to unattended speech [9, 21, 16]

Previous work by O'Sullivan et al. [21], Mirkovic et al. [16] used an approach called *stimulus-reconstruction*, which creates a linear mapping from the neural response to the speech stimulus. This takes into account delays between the speech stimulus and the neural response [8, 21, 16]. This approach is used in this study.

Revealing this connection between the envelope of speech and the neural response has opened up the possibility of applying other speech processing techniques to the problem of determining listening attention. Cepstral analysis is very useful for speech processing and so it is natural that it would be considered for this problem. There is currently no published research on the application of cepstral analysis to find connections between speech and the neural response. Therefore this project was intended as a preliminary investigation into whether or not cepstral processing could be used to determine listening attention.

Cepstral processing was first described in a paper published in 1963 by Bogert et al. [3], where it was observed that the logarithm of the power spectrum of a signal with an echo consisted of the power spectrum of the signal plus a periodic component due to the echo. Thus the cepstrum was put forward as a way of detecting echoes in a signal.

The significance of cepstral analysis in speech processing is a result of the way in which speech is produced. Humans and most other mammals vocalise in the same way, by pushing air through their larynx causing their vocal folds to vibrate. The sound produced is then filtered through the vocal tract giving the sound a characteristic "formant" structure [27]. When the vocal folds vibrate, *voiced* speech is created, otherwise it is *unvoiced*. The opening between the vocal folds is called the glottis. Vibrations created as air is pushed through the glottis is known as the *glottal impulse*. Thus, over short-time intervals, speech can be described as a convolution of a vocal tract response (determined by the shape of the vocal tract), and the glottal impulse (a train of impulses occuring at the frequency of oscillations of the vocal folds, know as the pitch period). That is, over short-time intervals, voiced speech can be described as an a series of echoes of the vocal tract response occurring at the rate of, and in response to, the glottal impulses.

This distinction between voiced and unvoiced speech led Michael Noll,in 1964, to apply cepstral analysis techniques to speech signals to short-time segments of speech to determine whether speech was voiced or unvoiced, and to detect the pitch period [18, 17].

In the early 1960s, at the same time as this early research into cepstral analysis, Al Oppenheim was researching a theory in non-linear systems theory called *homomorphic systems*. This theory proposed the idea that certain operations of signal combination, including convolution and multiplication, could be converted into a linear system, where conventional analysis techniques are well understood [28]. Similarities between the cepstrum and homomorphic systems led Oppenheim, Schafer, and Stockham to unite the two in a paper in 1968 [20] where the cepstrum was defined more clearly in the context of homomorphic systems.

The uniting of the theories of cepstral analysis and homomorphic systems was of great significance to speech processing. Now, as well as distinguishing between voiced and unvoiced speech and detecting pitch period, it was shown that it was possible to filter, or *lifter*, the cepstrum and then, using the *inverse* of the cepstrum, convert these separated parts of the cepstrum back into the original domain, giving the separated vocal tract impulse response and the glottal impulse train respectively. Thus, refining the definition of the cepstrum in the context of homomorphic systems led to the use of cepstral analysis in speech modelling and reconstruction. In this project we are interested in cepstral processing not for speech reconstruction purposes, but for finding connections between speech and EEG signals. Thus, it is not necessary for us to reconstruct speech but only to look for similarities between the cepstrums of the speech and EEG signals respectively, and so we work exclusively with the cepstrum and not with its inverse. The cepstrum is defined here in the context of homomorphic systems to provide the broader context for the reader.

This is a preliminary investigation into the possibilities of using a BCI system to determine listening attention for use in future hearing technology. Therefore the experiment conducted to collect the data used in this study is a very abstract scenario. Real-life scenarios are likely to bring a multitude of other complexities to this problem, however, at this early stage of research simplistic scenarios are essential in order to break the problem into more manageable steps. The data used in this study is described in detail in Appendix A. This data was also used in studies by O'Sullivan et al. [21], Power et al. [23]. Subjects undertook 30 trials, each of approximately 1 minute in length, where they were presented with 2 classic works of fiction: one to the left ear, and the other to the right ear. Each story was read by a different male speaker. Subjects were divided into two groups of 20 with each group instructed to attend to the story in either the left or right ear throughout all 30 trials.

As mentioned above, having found a connection between the envelope of speech and the EEG response, researchers are now considering other speech processing techniques for this challenge of determining listener attention using EEG. The usefulness of the cepstrum in relation to speech processing suggests that this could be a viable method. In this project we build upon the work of O'Sullivan et al. [21] who, as described above, were able to determine listener attention using the relationship between the respective attended and unattended speech signal envelopes, and the EEG response. Our research incorporates cepstral analysis into the stimulus-reconstruction model, comparing the cepstral coefficients of the respective attended and unattended speech streams and the EEG signal. Our intention was to see if the O'Sullivan et al. [21] results could be replicated or even improved using cepstral analysis.

The model developed in this project involved breaking corresponding speech and EEG signals into short-time segments over which cepstral coefficients were calculated. These coefficients were then used as stimuli in a stimulus-reconstruction model to predict the cepstral coefficients of an *attended* speech stream using the EEG signal. This prediction was then compared to the sets of cepstral coefficients for two respective speech streams and the set which most closely matched the predictions were determined to belong to the attended speech stream. With this method we were able to obtain high classification rates of over 90% suggesting that the method is effective at distinguishing between attended and unattended speech. However, since the model was trained and tested on only one dataset, it follows that the results are not conclusive at this stage. Although further tests on different datasets will be needed to confirm the validity of the model, the results obtained here suggest that the model has potential.

This thesis is divided into three sections: theory and methods, results, and conclusions. In chapters 2 to 4 we cover the theory and methods used in the project, in chapters 5 and 6 we present our results, and in chapter 7 we give conclusions and suggest possible further developments.

In chapter 2 we introduce speech and EEG signals and outline the fundamental digital signals processing techniques used. Cepstral processing is covered in chapter 3 and includes a short background of the cepstrum and its development in terms of homomorphic systems. Chapter 3 also explains in more detail how speech is produced and the relevance of the cepstrum in speech processing. Chapter 4 outlines the stimulus-reconstruction model. In the stimulus-reconstruction model the stimuli and responses can be determined arbitrarily. In our case the cepstral coefficients of the speech are taken as the stimulus and the cepstral coefficients of the EEG are taken as the response. Therefore it seemed appropriate to define the stimulus-reconstruction model in the context of the model used here. A more general definition can be found in [8].

Chapters 5 and 6 present the results obtained in our investigation. To keep our analysis simple to begin with, we started by creating simulations of the problem with varying levels of complexity. Chapter 5 gives an account of these first experiments where cepstral analysis was applied to simulations of the cocktail party problem. Following the simulations we moved on to real data. The results obtained when applying the stimulus-reconstruction model to real data are described in chapter 6. Matlab version R2017b was used throughout the project.

Chapter 7 concludes the project with an overview of the methods used and the results obtained. Some suggestions for future validations and improvements of the model are provided.

Chapter 2

Digital Signals Processing

This chapter introduces the signals and digital signals processing techniques that we worked with in this project, and provides definitions of mathematical concepts required to understand the chapters that follow. We begin with an outline of speech and EEG signals and their relation with respect to the research question we are considering. We then proceed to give a summary of methods used extensively throughout digital signals processing, including definitions of stationary stochastic processes, spectral densities, and non-parametric estimation methods of spectral densities.

2.1 Speech and EEG Signals

Sound is a pressure wave that passes through gases or liquids. With the application of sound energy, molecules alternate between compression and rarefaction along the path of the energy source. This is often represented by the graph of a sine wave, as shown in Figure 2.1. Huang et al. [11] and this representation of sound as an *acoustic wave* is how speech is usually represented. Although this description is entirely accurate, as Jan Schnupp eloquently explains in his book *Auditory Neuroscience, Making Sense of Sound*, sounds are so much more to us than this. Sounds provide us with valuable information about the physical properties of objects and events around us. In Schnupps words "*things make sounds, and different things make different sounds*". This is what we make use of, and capture what listening is truly about; learning about the objects and events that surround us. Schnupp also gives an interesting perspective on speech as effectively a form of telepathy in the sense that we beam our thoughts into another persons head using invisible vibrations [27].

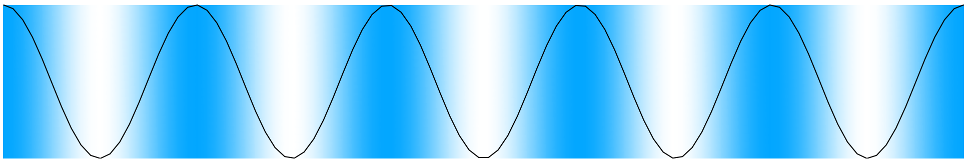


Figure 2.1 Sounds are alternating compressions and rarefactions in air molecules. A simple sound wave can be described by a sine wave. The blue colouring highlights the moments of greatest compression, and the white represents the moments of greatest rarefaction [11].

Even though speech and sounds mean more to us than a measure of vibrations in air molecules, this is what we measure when making sound recordings. To be useful in modern signal processing we need discrete rather than continuous recordings as they can be easily stored and manipulated by computers.

Mathematically, we can define an *analog signal* as a function that varies in continuous time, $x(t)$. Sampling this continuous signal, x, with a sampling period T, we obtain a discrete-time signal, or *digital signal*, which can be defined as $x[n] = x(nT)$. Note the use of square brackets and the time-index n for the digital signal, as opposed to the curved brackets and t used for the analog signal. This convention is used throughout this text to distinguish between discrete and continuous signals. The sampling frequency can be defined as $F = 1/T$. Throughout this text we will be working with Hertz (Hz), where 1Hz corresponds to a sampling frequency of 1 sample per second, and a sampling period of 1 second.

An EEG signal is a measurement of electrical currents that flow during synaptic excitations in the cerebral cortex. These electrical currents generate an electric field over the scalp that can be measured using EEG systems. EEG signals are recorded using multiple-electrodes placed either inside the brain (electrocorticogram (ECoG)), over the cortex under the skull (intracranial (iEEG) signals), or certain locations over the scalp (EEG) [26]. Scalp EEG is non-invasive and is the type of EEG signal used in this study. Any further reference to *EEG* is referring to scalp EEG signals. The non-invasive nature of EEG makes it particularly well suited to this problem as the BCI system that is the goal of this research will be most effective if it can be worn easily and with minimal effort.

Neural signals effectively range from 0.5-50Hz. Based on their frequency ranges, they have been grouped into five major categories: Delta (0.5-4Hz), Theta (4-8Hz), Alpha (8-12Hz), Beta (12-30Hz), and Gamma (above 30Hz, mainly up to 45Hz). In some studies the neural signal is filtered to consider only specific ranges [26]. However in this study we retain all frequencies of the EEG.

Similarly to speech, digital signals processing plays a fundamental role in EEG signals processing. However, EEG signals have lower frequencies than speech and so it follows that the sampling rate used can be lower whilst retaining the information of the signal. Therefore EEG is commonly recorded with a sampling rate of around 512Hz, whereas speech recorders commonly sample at a rate of 8-11kHz. This creates a challenge when working with speech and EEG as that they are most often sampled at different frequencies so adjustments have to be made before the signals can be compared. The sampling frequency of a signal can be reduced by a factor of k by taking every k-th value of a signal, a technique known as *downsampling*. The sampling frequency can be increased by a factor of k by interpolating between points in a signal with $k-1$ values.

2.2 Stationary Stochastic Processes

Let $x(t)$ denote a continuous-time signal (an analog signal). By sampling this signal x with a sampling time interval T_s, that is $t = nT_s$, we obtain a discrete-time data sequence defined as $\{x[n]; n = 0, \pm 1, \pm 2, ...\}$, also known as a digital signal. Let this discrete-time data sequence be a stochastic process. From the definition given by Lindgren et al. [13] this stochastic process has first and second moments defined as

$$\begin{aligned}
m(n) &= \mathbb{E}[x[n]] && mean\ value\ function \\
v(n) &= \mathbb{V}[x[n]] && variance\ function \\
r(m,n) &= \mathbb{C}[x[m]x[n]] && covariance\ function \\
b(m,n) &= \mathbb{E}[x[m]x[n]] && second-moment\ function \\
\rho(m,n) &= \rho[x[m]x[n]] && correlation\ function
\end{aligned}$$

In this project, we will be working with stationary stochastic processes, in specific with *weakly stationary* processes. Such processes are those that even after a change or displacement of the time scale their statistical properties (first and second order moments) remain the same [13]. A discrete-time data sequence, $\{x[n]\}$, is a *weakly stationary* process if it has constant mean $m(n)$ and its covariance function $r(m,n) < \infty$ is finite and is dependent only in the time changes $\tau = n - m$ [13].

Let E_s denote the energy of the discrete signal $x[n]$, if this sequence $x[n]$ has finite energy, i.e.

$$E_s = \sum_{n=-\infty}^{\infty} |x[n]|^2 < \infty, \tag{2.1}$$

then the sequence $x[n]$ has a discrete-time Fourier Transform (DTFT) defined as

$$X(f) = \sum_{n=-\infty}^{\infty} x[n]e^{-i2\pi fn}, \tag{2.2}$$

for frequency f with period 2π [29]. For this project, we will be working in the frequency domain, f. Note for reference that in some text notation can be found as $\omega = 2\pi f$.

The original sequence $x[n]$ is then obtained through the corresponding inverse discrete-time Fourier Transform (IDTFT)

$$x[n] = \int_{-\frac{1}{2}}^{\frac{1}{2}} X(f)e^{i2\pi fn}df, \tag{2.3}$$

and the Energy Spectral Density is defined as

$$S(f) = X(f)X^*(f) = |X(f)|^2, \tag{2.4}$$

where $(\cdot)^*$ denotes the complex-conjugate of a scalar variable or the conjugate transpose of a vector or matrix [29].

Speech and EEG signals are generally non-stationary, however, they are quasi-stationary, that is, they can be considered stationary within short time intervals [26]. One of the fundamental assumptions made in speech processing is that, when considered over short time intervals (generally 20-25ms), speech signals can be considered stationary [2].

2.3 Digital Systems

Huang et al. [11], refer to digital systems as those that, given an input digital signal $x[n]$, can generate an output signal $y[n]$:

$$y[n] = T\{x[n]\}. \tag{2.5}$$

In general, a digital system T is defined to be linear if and only if

$$T\{\alpha_1 x_1[n] + \alpha_2 x_2[n]\} = \alpha_1 T\{x_1[n]\} + \alpha_2 T\{x_2[n]\} \ \forall \alpha_1,\ \alpha_2 \in \mathbb{R}, \tag{2.6}$$

for any signals $x_1(t)$ and $x_2(t)$. T is defined to be time-invariant if the output is not affected by the particular point in time at which the input is applied to the system. According to

Huang et al. [11], linear time-invariant systems can be described by

$$x[n] = \sum_{k=-\infty}^{\infty} x[k]h[n-k] = x[n] * h[n].$$

By substituting $x[n] = e^{i2\pi fn}$ in the previous equation we have

$$y[n] = \sum_{k=-\infty}^{\infty} h[k]e^{i2\pi f[n-k]},$$

and using the commutative property of the convolution we obtain the following expression

$$\begin{aligned} y[n] &= \sum_{k=-\infty}^{\infty} h[k]e^{i2\pi f[n-k]} \\ &= \sum_{k=-\infty}^{\infty} h[k]e^{i2\pi fn}e^{-i2\pi fk} \\ &= e^{i2\pi fn} \sum_{k=-\infty}^{\infty} h[n]e^{-i2\pi fk} \\ &= e^{i2\pi fn}H(f), \end{aligned}$$

where $H(f)$ is the discrete-time Fourier Transform of $h[n]$ and is expressed as a function of the frequency with period 2π. It is called the system's frequency response or transfer function [11]. The impulse response is accordingly defined by,

$$h[n] = \int_{-\frac{1}{2}}^{\frac{1}{2}} H(f)e^{i2\pi fn}df,$$

the corresponding inverse discrete-time Fourier transform [11].

2.4 Power Spectral Density

According to Stoica and Moses [29], if a signal has finite average power, then it can be described by the average power spectral density. Throughout this project we will refer to the average spectral density of a signal as the Power Spectral Density (PSD). The PSD is a way to characterize and provide details about how this power is spread among frequencies.

We will use the definition of covariance function given by Lindgren et al. [13]. For $m < n$, let τ be the time lag given by $\tau = n - m$, then the covariance function is given by

$$r(\tau) = r(m,n) = C[x[m], x[n]] = \mathbb{E}(x[m]x[n]) - m_x(n)m_x(m), \tag{2.7}$$

where $m_x(n)$ and $m_x(m)$ are the mean value functions of $x[n]$ and $x[m]$ respectively. For a discrete-time signal $x[n]$ with zero mean it follows that

$$r(\tau) = \mathbb{E}(x[n]x[n-\tau]). \tag{2.8}$$

We can find an asymptotically unbiased estimator for the covariance function using its corresponding sample covariance and assuming that the mean of the process $m(n) = m$ is known [13]. The following theorem formulates such an estimator. Its proof can be found in [13].

Theorem 1 *The estimator $\hat{r}(\tau)$ of the covariance function $r(\tau)$ is given by*

$$\hat{r}(\tau) = \frac{1}{N} \sum_{n=1}^{N-|\tau|} (x[n] - m)(x[n+|\tau|] - m), \tag{2.9}$$

and it is asymptotically unbiased if

$$\lim_{N \to \infty} \mathbb{E}[\hat{r}(\tau)] \to r(\tau).$$

Stoica and Moses [29] used eq. (2.8) to define the PSD (viewed as a function of frequency) as the DTFT of the covariance function:

$$R(f) = \sum_{\tau=-\infty}^{\infty} r(\tau) e^{-i2\pi f \tau}. \tag{2.10}$$

We can obtain back the covariance function $r(\tau)$ via the Inverse DTFT of $R(f)$ [29]:

$$\begin{aligned}
IDFT\{R(f)\} &= \int_{-1/2}^{1/2} R(f) e^{i2\pi f \tau} df && (2.11)\\
&= \int_{-1/2}^{1/2} \sum_{k=-\infty}^{\infty} r(k) e^{-i2\pi f k} e^{i2\pi f \tau} df && (2.12)\\
&= \sum_{k=-\infty}^{\infty} r(k) \int_{-1/2}^{1/2} e^{-i2\pi f (k-\tau)} df && (2.13)\\
&= r(\tau). && (2.14)
\end{aligned}$$

Now that we introduced the previous definitions for PSD, we can now mention the nonparametric methods used in this project to obtain an estimate of the PSD, $R\hat{(f)}$.

2.5 Non-parametric Methods of Spectral Estimation

In this section, we introduce the non-parametric methods of spectral estimation that were considered throughout this project, as well as some of their properties. When working with non-parametric methods no assumptions are made about the underlying distribution of the data. When PSD is estimated using non-parametric methods, there is a trade-off between resolution in the spectrum and high variance. The periodogram is known for providing a good resolution in the peaks of the spectrum, but it has high variance [29]. Numerous modified methods, for example the ones introduced by Bartlett [1] and Welch [31], have been created with the aim of reducing this high variance (characteristic of the periodogram), but this variance reduction comes with a loss of resolution at the peaks [29]. For this project, we will limit ourselves to using the periodogram, the periodogram with Hanning windowing, the Welch method, and multitaper methods.

2.5.1 Periodogram

Suppose we are interested in estimating the spectral density of a stationary process $X_n; n \in \mathbb{Z}$, which has been sampled to obtain a real value data sequence $\{x([n]; n = 0, \pm1, \pm2, ...\}$ and fulfils eq. (2.14). We can make a spectral estimation of this data sequence using the periodogram, defined as

$$\hat{R}(f) = \frac{1}{N} \left| \sum_{n=0}^{N-1} x[n] e^{-i2\pi f n} \right|^2 , \tag{2.15}$$

we can use the conjugate of the term inside the absolute value to separate the power and expressed eq. (2.15) as

$$\begin{aligned} \hat{R}(f) =& \frac{1}{N} \left[\sum_{m=0}^{M-1} x[m] e^{-i2\pi f m} \right] \left[\sum_{n=0}^{N-1} x[n] e^{i2\pi f n} \right] && (2.16) \\ =& \frac{1}{N} \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} x[m] x[n] e^{-i2\pi f (m-n)}. && (2.17) \end{aligned}$$

Let $\tau = m - n$. If we first sum over all possible M values and collect all terms $x[m]x[n]$, we can see that for $\tau = 0 \iff m = n$, we will end up with N possible terms

$$x[0]x[0],\ x[1]x[1],\ ...,\ x[N-1]x[N-1].$$

For $\tau = 1 \iff m = n+1$, we will end up with $N-1$ possible terms

$$x[0]x[1],\ x[1]x[2],\ ...,\ x[N-1]x[N],$$

and so on for values of m up to $N-1-|\tau|$ as $\tau \in [-N+1, N+1]$. We see that all possible combinations can be expressed as $x[n]x[n-|\tau|]$ and therefore write eq. (2.15) as

$$\hat{R}(f) = \frac{1}{N} \sum_{\tau=-N+1}^{N-1} \sum_{n=0}^{N-1-|\tau|} x[n]x[n+|\tau|]e^{-i2\pi f\tau}. \tag{2.18}$$

Assuming that the mean of the process $m(n) = m$ is known and equal to zero, we can recognize the expression for the estimator $\hat{r}(\tau)$ from eq. (2.9) in order to express the periodogram in terms of the covariance function as

$$\hat{R}(f) = \sum_{\tau=-N+1}^{N-1} \hat{r}(\tau)e^{-i2\pi f\tau}. \tag{2.19}$$

We can compute the expected value of the periodogram from this eq. (2.19) using the expression from eq. (2.18) for $\hat{r}(\tau)$, as it is shown in [13].

$$\mathbb{E}[\hat{R}(f)] = \sum_{\tau=-N+1}^{N-1} \mathbb{E}[\hat{r}(\tau)]e^{-i2\pi f\tau} \tag{2.20}$$

here, if we expand the term $\mathbb{E}[\hat{r}(\tau)]$ by letting $N \to \infty$ we obtain the asymptotically unbiased estimator for $r(\tau)$ yielding the next expression

$$\mathbb{E}[\hat{r}(\tau)] = \frac{1}{N} \sum_{n=1}^{N-|\tau|} \mathbb{E}[(x[n]-m)(x[n+|\tau|]-m)] \tag{2.21}$$

$$= \frac{1}{N} \sum_{n=1}^{N-|\tau|} r(\tau) \tag{2.22}$$

$$= \left(\frac{N-|\tau|}{N}\right) r(\tau) \tag{2.23}$$

$$= \left(1 - \frac{|\tau|}{N}\right) r(\tau). \tag{2.24}$$

Substituting eq. (2.24) in eq. (2.20) we can express the expected value of the periodogram as

$$\mathbb{E}[\hat{R}(f)] = \sum_{\tau=-N+1}^{N-1} \mathbb{E}[\hat{r}(\tau)]e^{-i2\pi f\tau} \tag{2.25}$$

$$= \sum_{\tau=-N+1}^{N-1} \left(1 - \frac{|\tau|}{N}\right) r(\tau)e^{-i2\pi f\tau}. \tag{2.26}$$

From this expression we obtain the so-called *lag window* $w[n] = \left(1 - \frac{|\tau|}{N}\right)$. When $N \to \infty$, $\hat{R}(f)$ yields an estimate of the PSD which is asymptotically unbiased [13]. But when dealing with values of $N < \infty$, i.e., when working with finite length sequences of data that were obtained from infinite length sequence of data; we will run into spectrum bias, $B(f) = \mathbb{E}[\hat{R}(f)] - R(f)$. Figure 2.2 shows an example of an estimate of the PSD using periodogram for a signal, $x[n]$ with main frequency located at 200 Hz. We can note the presence of side-lobes located next to main peak corresponding to the main frequency.

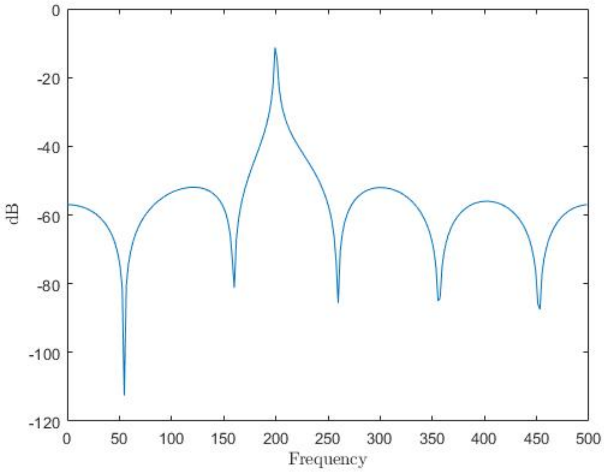


Figure 2.2 Periodogram of a sinusoidal sequence with main frequency of 200 Hz. The main peak at 200 Hz correspond to this main frequency. The side-lobes are inconsistencies that occur when using the periodogram for PSD estimation.

On the one hand, we have this bias that arises when working with finite length sequences of data. On the other hand, we have the characteristic variability in PSD estimations which is not reduced even for large values of N. These two reasons make from the periodogram an inconsistent estimator of the PSD [13]. As an alternative to deal with this situation, one can try different types of *lag window*. The *windows* that were used within this project will be presented in the next section.

2.5.2 Windowing

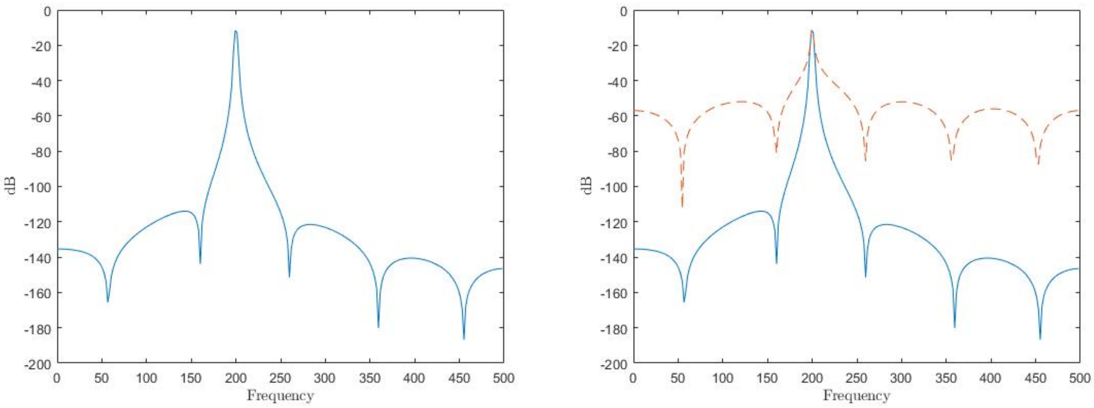
When computing the periodogram of a signal, one might notice the presence of some side-lobes next to the peaks that are present at the main frequencies in the spectrum. If the power that corresponds to main frequencies is leaked to this side-lobes, it is possible to misinterpret where the main frequencies of the signal are located. In order to gain a better resolution in the peaks and reduce the side-lobes and reduce this bias a common practice is to *window* data. The use of *lag window* functions w lead to a modified version of the periodogram [13], which can be derived from eq. (2.26)

$$\hat{R}_w(f) = \frac{1}{n}\left|\sum_{n=0}^{N-1} x[n]w[n]e^{-i2\pi fn}\right|^2. \tag{2.27}$$

Multiple windows have been developed for PSD estimation. One of the most relevant and widely used is the Hanning window [13],

$$h(t) = \frac{1}{2} - \frac{1}{2}\cos\frac{2\pi t}{n-1}, t = 0, ..., n-1. \tag{2.28}$$

This was the selected window used in this project. For a sequence $x[n]$, if we use a Hanning window in the periodogram, the side-lobes will drop more rapidly, stopping the power from leaking to the sides and instead remaining around the main lobes. The downside of this is that the main-lobe in the Hanning windowed periodogram is wider than the standard periodogram. This wider main-lobe reduces the resolution and means that if, for example, two frequency peaks are close together, they may appear as a single peak [13]. This is illustrated in Figure 2.3 where we can see an estimate of the PSD for the same sequence $x[n]$ used for the example in the previous section 2.5.1. Figure 2.3(a) shows the periodogram using a Hanning window and Figure 2.3(b) shows both periodograms using a Hanning window and without any window function (red dashed line). In Figure 2.3(b) we can see the effect of the side-lobes.



(a)

(b)

Figure 2.3 Figure (a) shows the PSD estimate of the signal $x[n]$ estimated with a modified version of the periodogram using a Hanning window. Figure (b) shows this same PSD estimate and the red dashed line corresponds to the PSD estimate using periodogram without any window function.

2.5.3 The Welch Method

This method is also known as the weighted overlap segmented averaging (WOSA) method and was first introduced by P. Welch in 1967 [31]. It is a nonparametric method for estimating the PSD, modified from a method introduced by Bartlett [1] in 1948. Both methods were developed with the aim of reducing the variance in the periodogram. In order to describe the Welch method, we need first to present the Bartlett method which can be summarized in 3 steps [24]. For a given finite sequence. $x[n]$:

1. Divide n data points into K non-overlapping segments of length M with the $n-$th point of the $j-$th segment denoted by

$$x_j[n] = x[n + jM], \text{ for } j = 0, 1, \ldots, K-1 \text{ and } n = 0, 1, \ldots, M-1. \tag{2.29}$$

2. Compute the periodogram using eq. (2.15) for each $j-$th segment. This means that we will obtain K periodograms

$$\hat{R}(f) = \frac{1}{M}\left|\sum_{n=0}^{M-1} x_j[n] e^{-i2\pi f n}\right|^2, \text{ for } j = 0, 1, \ldots, K-1. \tag{2.30}$$

3. The Bartlett PSD estimate is given by the average of the periodograms for the K segments

$$\hat{R}^{(B)}(f) = \frac{1}{K}\sum_{j=0}^{K-1} R^{j}(f) \tag{2.31}$$

Now that we have given an outline of the Bartlett method we can describe the Welch method. The Welch [31] method consists of averaging modified periodograms using overlapping segments. Let now the K segments to overlap

$$x_j[n] = x[n + jD], \text{ for } j = 0, 1, \ ..., L-1 \text{ and } n = 0, 1, \ ..., \ M-1, \tag{2.32}$$

where jD is the point where the $j-$th sequence begins. The Welch method applies the window function w (see eq. (2.27)) to the sequence $x[n]$ before computing the periodogram. The result of this windowing is a modified periodogram

$$\hat{R}^{(j)}(f) = \frac{1}{MU}\left|\sum_{n=0}^{M-1} x_j[n]w[n]e^{-i2\pi fn}\right|^2, \text{ for } j = 0, 1, \ ..., \ L-1. \tag{2.33}$$

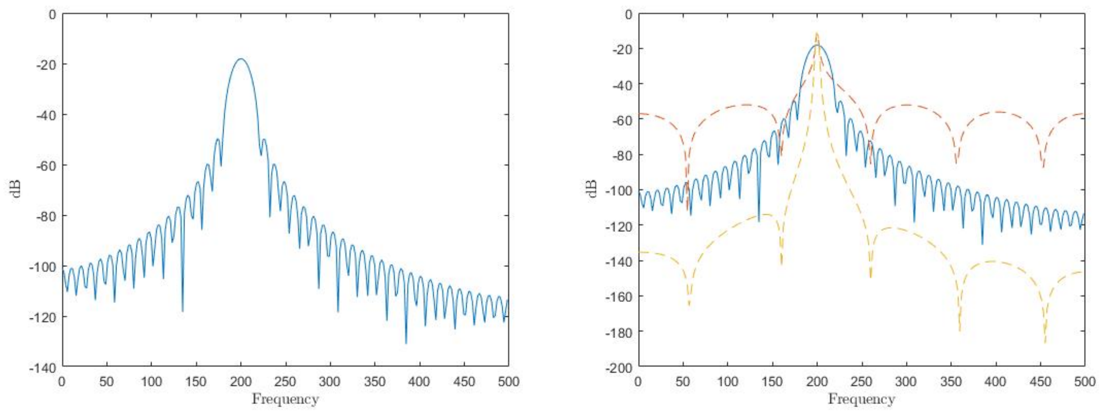
This modified periodogram needs to be normalized [24]. This is done using a normalizing factor U for the power in the function window w

$$U = \frac{1}{M}\sum_{n=0}^{M-1} w^2[n]. \tag{2.34}$$

The PSD estimate method using the Welch [31] method is given by the average of these modified periodograms

$$\hat{R}^{(W)}(f) = \frac{1}{L}\sum_{j=0}^{L-1} R^{(j)}(f). \tag{2.35}$$

For the same sequence $x[n]$ used in section 2.5.1, an illustration of the PSD estimate using the Welch method is shown in Figure 2.4. Figure 2.4(a) shows the PSD using Welch method and Figure 2.4(b) shows the previous PSD estimations found in this chapter superimposed over the one obtained using the Welch method. We see the reduction of the side-lobes but the main peak is wider.



(a)

(b)

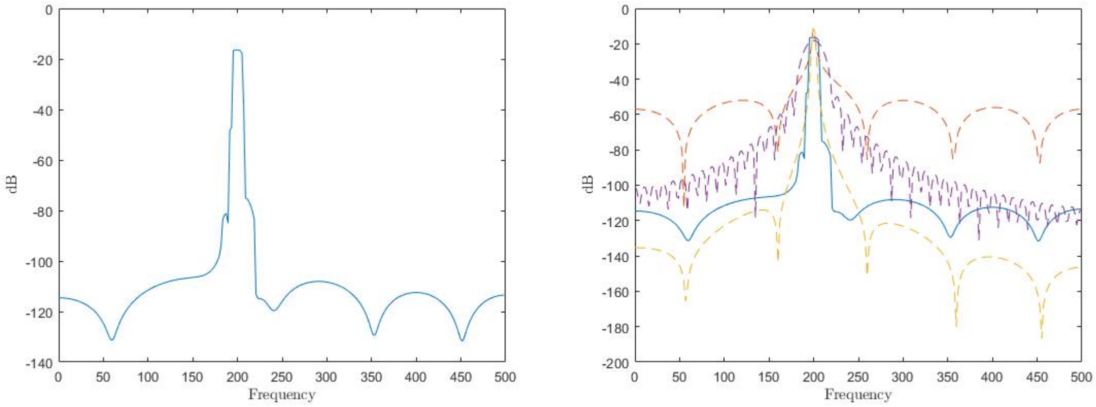
Figure 2.4 Figure (a) shows the PSD estimate of the signal $x[n]$ estimated using Welch method. Figure (b) shows the same PSD estimate with the previous PSD estimations found in this chapter. The red dashed line corresponds to the PSD estimate using periodogram without any window function. The yellow dashed line corresponds to the modified periodogram using a Hanning window.

2.5.4 Multitapers

As discussed in the previous section 2.5.3, a reduction of the variance and the side-lobes in the periodogram is accomplished when averaging over a certain number of periodograms. The inconvenience of this method is that the number of data points that are actually used for computing each periodogram is small compared to the total length of the sequence [24]. An alternative to this was introduced by Thomson [30]. His method consists of using multiple windows, obtaining multiple windowed PSD estimates and taking the average of these multiple estimates

$$\hat{R}^{(M)}(f) = \frac{1}{K}\sum_{k=0}^{K-1}\hat{R}^{(k)}(f) = \frac{1}{K}\left|\sum_{k=0}^{K-1}\sum_{n=0}^{N-1} x[n]w[n]e^{-i2\pi fn}\right|^2. \tag{2.36}$$

An illustration of this PSD estimation method is shown in Figure 2.5(a). Figure 2.5(b) gives a comparison of the multitaper method with the other methods discussed in this chapter.



(a) (b)

Figure 2.5 PSD estimation using multitapers is shown in Figure (a). In Figure (b) A comparison of this estimate with the other estimates obtained using the methods described throughout this chapter.

Chapter 3

Cepstral Processing

3.1 Cepstral Processing

In 1963, Bogert et al. [3] observed that the logarithm of the power spectrum of a signal plus it's echo (i.e. a signal followed by a delayed and scaled replica) consists of the logarithm of the signal spectrum and a periodic component due to the echo. With further spectral analysis they found that they could identify the periodic component in the log spectrum and therefore had a new indicator for the occurrence of an echo [2].

Applying spectral analysis techniques *on* the spectrum of a signal, Bogert et al. [3] came up with a new vocabulary to reflect this. They chose to rearrange the first syllable of words taken from spectral analysis in order to highlight the connections between the two, while also making clear the difference between the methods. Hence, working in the *quefrency domain*, the spectrum of the log spectrum of a time waveform came to be know as the *cepstrum*, and filtering of this cepstrum was named *liftering*. Harmonics are named *rahmonics* in the quefrency domain [19].

The original definition of the cepstrum was based on the power spectrum of an analog signal. However, the application of the cepstrum using modern computing techniques requires digital processing and thus a clear definition of the cepstrum in terms of discrete-time signal theory was required [2]. For discrete-time signals, the cepstrum is defined as the IDTFT of the natural logarithm of the DTFT of the signal. That is, for a discrete-time signal, $x[n]$, the discrete-time cepstrum is given by

$$c[n] = \int_{-\frac{1}{2}}^{\frac{1}{2}} \log|X(f)|e^{i2\pi fn}df, \tag{3.1}$$

3.2 Homomorphic Systems

In the early 1960s, at the same time as Bogert et al. were working on the theory of the cepstrum, Al Oppenheim was researching a new theory in nonlinear signals processing referred to as *homomorphic systems*. Oppenheims work was based on applying linear vector space theory to signals processing. The idea was that certain operations of signal combination (particularly convolution and multiplication) satisfy the same axioms as vector addition in linear vector space theory [2].

Of interest here is the class of homomorphic systems for convolution. This is represented through the diagram in Figure 3.1

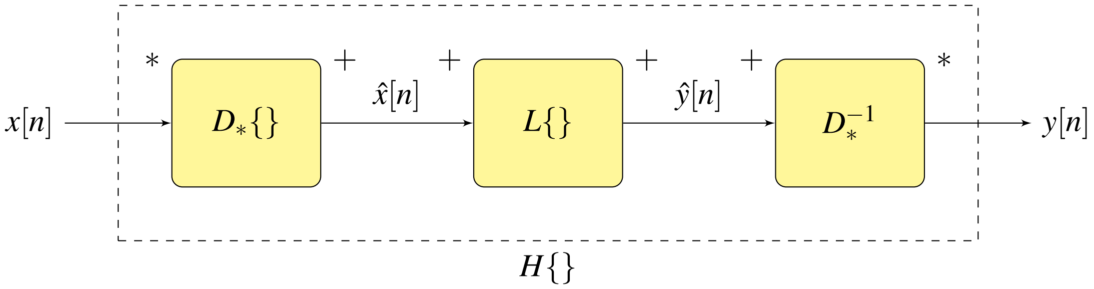


Figure 3.1 Homomorphic system for convolution

$D_*\{\ \}$ represents the *characteristic system for convolution* and transforms a combination by convolution into a corresponding combination by addition. $D_*\{\ \}$ is therefore defined by the property that when $x[n] = x_1[n] * x_2[n]$, the corresponding output is

$$\begin{aligned} \hat{x}[n] &= D_*\{x_1[n] * x_2[n]\} \\ &= D_*\{x_1[n]\} + D_*\{x_2[n]\} \\ &= \hat{x}_1[n] + \hat{x}_2[n] \end{aligned} \tag{3.2}$$

$L\{\ \}$ is an ordinary linear system that satisfies the principle of superposition with addition as the input and output operation for signal combination. The inverse characteristic system, $D_*^{-1}\{\ \}$, must transform a sum into a convolution. The operations that apply at each stage are written at the top corners of each block [2].

The calculation of the cepstrum gives a sequence of mathematical operations that satisfy the property of eq. (3.2). That is, we can represent $\hat{x}[n]$ by the equations

$$\hat{X}(f) = \log[X(f)] \tag{3.3}$$

$$\hat{x}[n] = \int_{-\frac{1}{2}}^{\frac{1}{2}} \hat{X}(f) e^{i2\pi f n} \mathrm{d}f \tag{3.4}$$

where $X(f)$ is the DTFT, eq. (3.3) is the *complex* logarithm, and eq. (3.4) is the IDTFT of the complex function $X(f)$. Note that in eq. (3.3), the complex logarithm is used, which is defined as

$$\log\{X(f)\} = \log|X(f)| + i \cdot \arg\left[X(f)\right]. \tag{3.5}$$

This sequence is illustrated using the diagram in Figure 3.2. The inverse of the characteristic system for convolution is shown in Figure 3.3, and inverts the effect of the logarithm by applying an exponential function [2].

The *cepstrum* is the *real* part of the complex cepstrum, and differs from the complex cepstrum only in the fact that the log of the magnitude of the spectrum is taken rather than the complex logarithm. The *real cepstrum* is the most widely used in speech technology [11] and as such we only consider the cepstrum in this project. Further references to the *cepstrum* are referring to the real cepstrum.

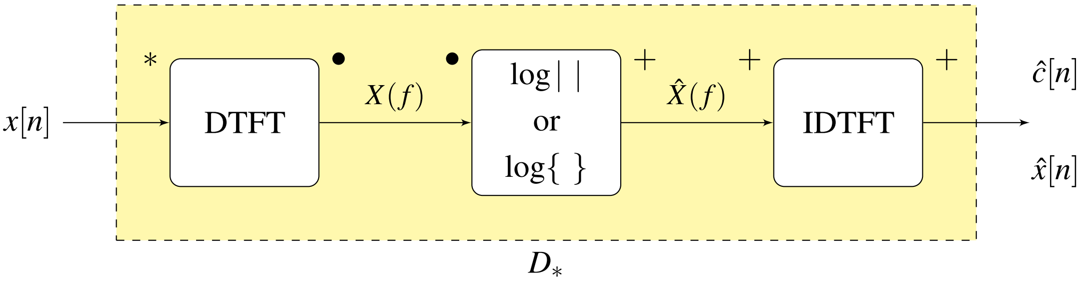


Figure 3.2 Characteristic system for convolution

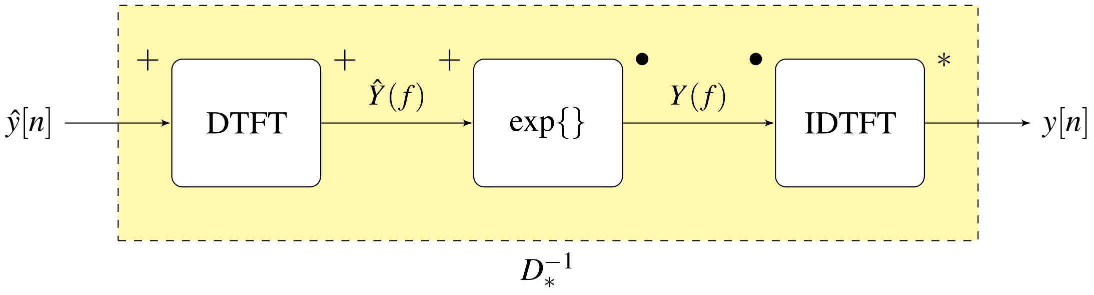


Figure 3.3 Inverse characteristic system for convolution

The connection between the characteristic system for convolution and the cepstrum was first made in 1965, during a discussion between Oppenheimer and Jim Flanagan, of Bell Telephone Laboratories. Flanagan remarked that the homomorphic system for convolution reminded him of the cepstrum proposed by Bogert et al. [3]. The characteristic systems for convolution given in Figures 3.2 and 3.3 were developed by *Oppenheim*, *Schafer*, and *Stockham*, and were first published by Oppenheim et al. [20]. Due to the similarities with the cepstrum described by Bogert et al. [3], when *Oppenheim*, *Schafer*, and *Stockham* published their work in 1968, they called the output of the characteristic system for convolution that uses the complex logarithm, (eqs. (3.3) and (3.4)) the *complex cepstrum*.

Homomorphic filtering, or *liftering*, is achieved by multiplying the complex cepstrum or cepstrum of a signal by a sequence $l[n]$, that is

$$y[n] = l[n]\hat{x}[n]. \tag{3.6}$$

This is one of the choices for the linear system, $L\{\}$ shown in Figure 3.1 [2]. It is using liftering that we can separate the convolution $x_1[n] * x_2[n]$, given in eq. (3.2), into the two respective components $x_1[n]$ and $x_2[n]$. This can be seen more clearly in the following example.

3.3 Cepstral Processing Example

Here we provide a example as an illustration of the cepstral processing techniques outlined above. Consider a signal with a simple echo, $x[n]$. We can write this as

$$x[n] = s[n] + \alpha s[n-\tau], \tag{3.7}$$

where τ and α represent the delay and scaling of the echo respectively.

The spectral density of eq. (3.7) is given by

$$|X(f)|^2 = |S(f)|^2 \left[1 + \alpha^2 + 2\alpha \cos(2\pi f \tau)\right] \tag{3.8}$$

Therefore, eq. (3.8) shows that the spectral density of a signal with an echo consists of the spectrum of the original signal modulating a periodic function of the frequency, f. Taking the logarithm of the spectrum, this product is converted to the sum of two components, that is

$$\log |X(f)|^2 = \log |S(f)|^2 + \log[1 + \alpha^2 + 2\alpha \cos(2\pi f \tau)] \tag{3.9}$$

As a waveform, eq. (3.9) has an additive periodic component with τ, the echo delay, as its "fundamental frequency". Taking the spectrum of this log spectrum would therefore show a peak where the original signal contained an echo [19].

Let $\tau = 100$, $\alpha = 0.8$, and $x_1[n]$ be the signal of length $N = 26$ shown in Figure 3.4. The signal first occurs at $n = 50$ and its echo occurs at $n = 150$. This can be considered in relation to eq. (3.2), where $x[n]$ is a convolution of $x_1[n]$ with $x_2[n]$. In this case, $x_2[n]$ is a signal with an impulse of amplitude 1 at time $n = 50$, and an impulse of amplitude α at time $n = 150$.

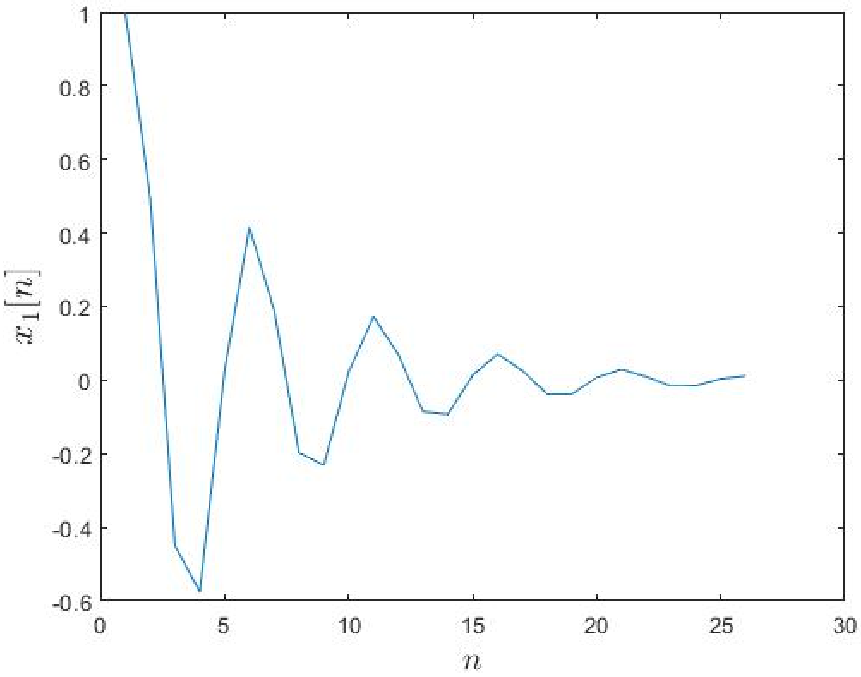


Figure 3.4 Signal $x_1[n]$, an impulse response of length N=26.

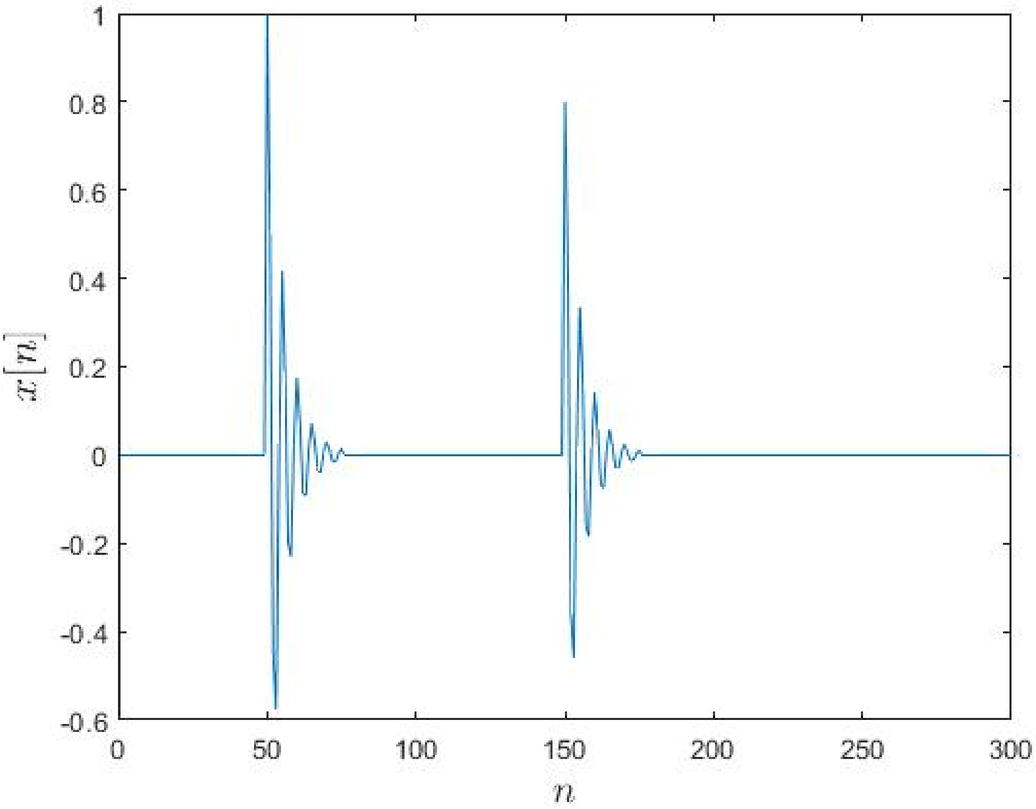


Figure 3.5 $x[n]$. This is signal $x_1[n]$ plus its echo (i.e. the impulse response $x_1[n]$ followed by a delayed and scaled replica). The signal occurs at $n = 50$ and its echo occurs at $n = 150$.

The PSD and the logarithm of the PSD of $x[n]$ are shown respectively in Figure 3.6. For this example the PSD of $x[n]$ was obtained using the periodogram method. Observing the spectral density and the log spectral density of $x[n]$ we can see that the spectrum peaks $f = 0.2$, indicating that this is the dominant frequency in the signal. We can also see that the spectrum appears to have a high frequency and a low frequency component to it. There is the slowly varying envelope which peaks at $f = 0.2$, and also a fast varying fluctuation with visible periodicity.

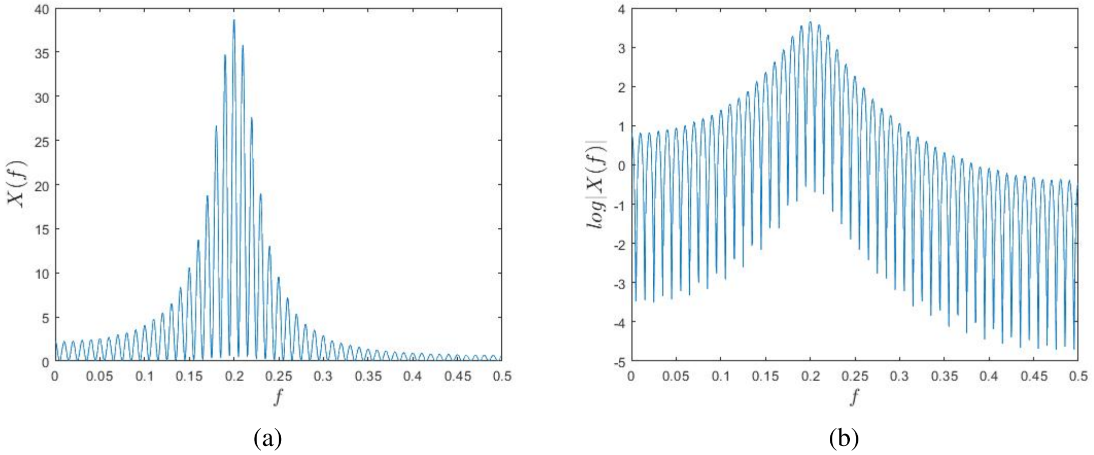


Figure 3.6 Figure (a) shows the Power Spectral Density of the signal $x[n]$ estimated using the periodogram. Figure (b) shows the logarithm of the absolute value of the Power Spectral Density estimate.

Applying the IDTFT to the log spectral density we obtain the cepstrum, shown in Figure 3.7. The sharp peak at $n = 50$ is the first *rahmonic* peak, and is followed by other *rahmonics* at multiples of 100. This corresponds to the delay of the echo in $x[n]$ being $\tau = 100$ time points after the initial signal.

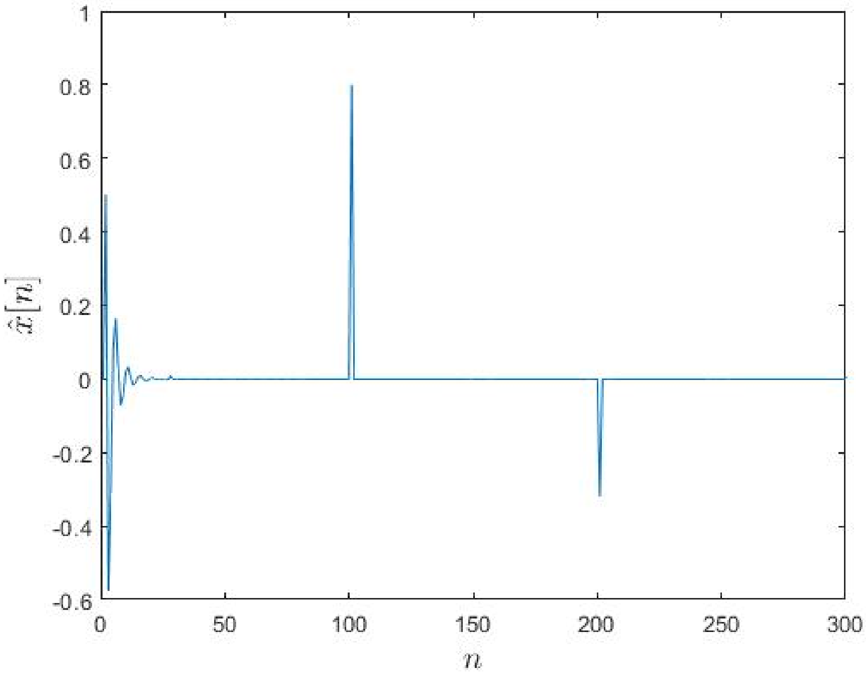


Figure 3.7 $\hat{x}[n]$, the cepstrum of $x[n]$

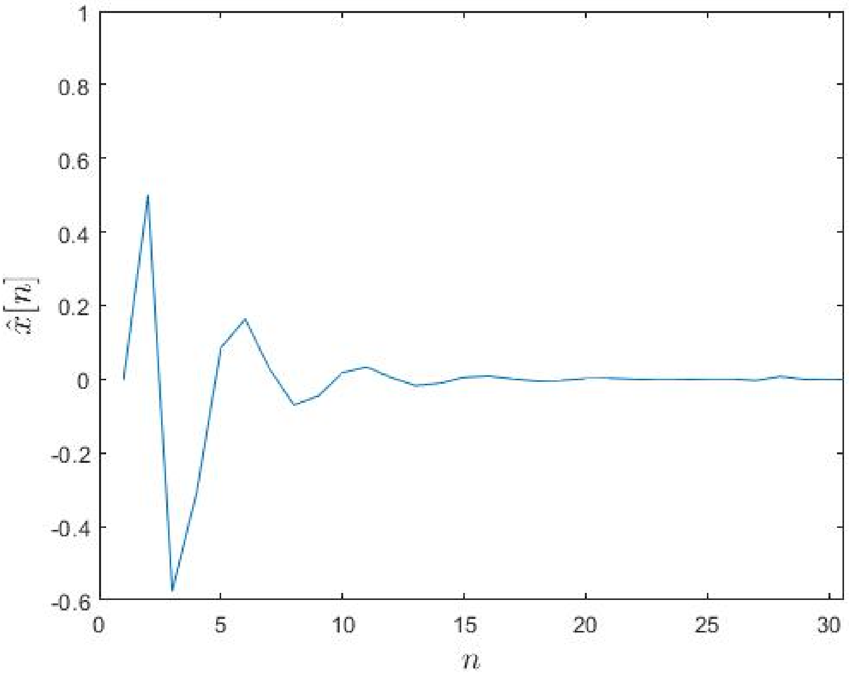


Figure 3.8 Low quefrency range of the cepstrum $\hat{x}[n]$.

The low quefrency region of the cepstrum corresponds to the slowly varying components of the log magnitude spectrum, while the high quefrency region corresponds to the rapidly varying components [2]. Making use of the concept introduced by Bogert et al. [3] we can separate these different components of the cepstrum.

Applying a low-pass lifter to the cepstrum we obtain the low-quefrency component which is shown in Figure 3.9 with the original impulse response, $x_1[n]$.

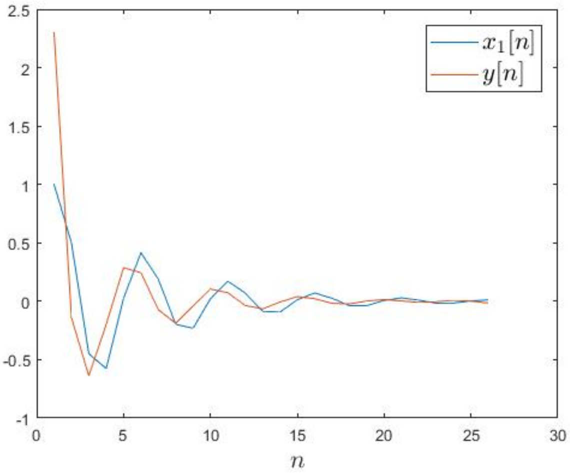


Figure 3.9 The impulse response, $x_1[n]$, and the reconstruction, $y[n]$, obtained by deconvolving $x[n]$ using cepstral analysis.

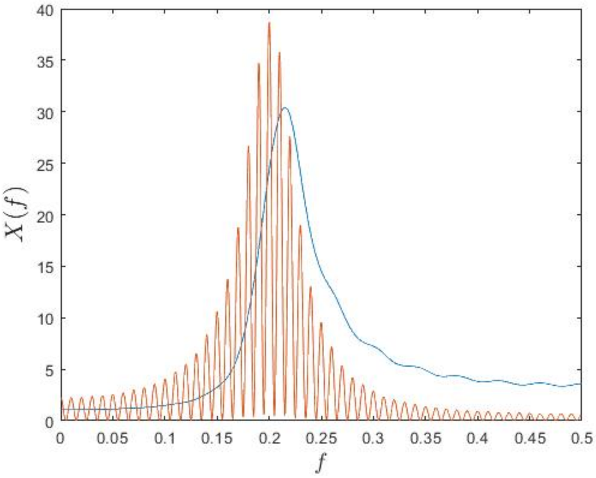


Figure 3.10 The spectral density of $x[n]$ (red) plotted together with the spectral density of the spectral density of $y[n]$, obtained by deconvolving $x[n]$ using cepstral analysis.

3.4 Cepstral Processing of Speech

Humans create speech signals through a series of controlled movements of their lungs, vocal cords, tongue, and lips [2]. Speech can be separated into two sound types, *voiced* and *unvoiced*. Voiced speech has a roughly regular pattern in its time-frequency structure whereas unvoiced speech does not. Voiced sounds also typically have more energy. Voiced speech is created when the vocal folds vibrate as a phoneme is articulated, otherwise the speech is unvoiced. To oscillate, the vocal folds are brought close enough together so that air pressure builds up beneath the larynx. As air pressure builds eventually the folds are forced apart but, due to the elasticity of the muscles and tissues of the vocal folds and larynx, they soon come together again when the air pressure equalises. These successive bursts of air create voiced sounds. The size and stiffness of the vocal folds, and the amount of air pressure below the vocal folds determines the rate at which they open and close. A speaker can raise and lower the pitch of a voiced sound by altering these factors. The rate of cycling (opening and closing) of the vocal folds is known as the *fundamental frequency*, or *pitch period* (P_0). Pitch periods vary from around 60 Hz for a large man, to 300 Hz or higher for a small woman or child [11].

A simple discrete-time model for a speech signal which mimics the way voiced and unvoiced speech is produced is given in Figure 3.11. The *impulse train generator* models the glottal pulse excitation (corresponding to voiced speech), and the *Random noise generator* models the fricative excitation (corresponding to unvoiced speech) of the vocal tract. The time-varying digital filter is a linear system with a slowly time-varying impulse response which models the frequency resonances (formants) of the human vocal tract, the *vocal tract response* [2].

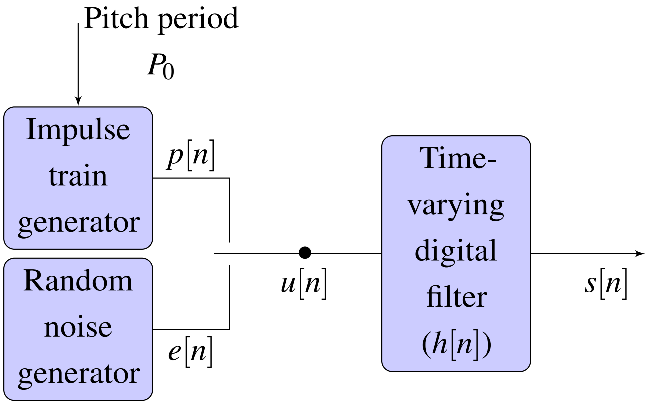


Figure 3.11 Discrete-time model for a speech signal

One of the main assumptions in speech processing is that, taken over short-time intervals, or *frames* (most commonly 20-30ms), speech signals are stationary (see section 2.2). A further assumption is that speech properties such as pitch period and vocal tract response are constant over these frames [2].

As such, over a frame of length L, we assume that a speech signal $x[n]$ can be modeled as a convolution of the excitation $u[n]$ (voiced ($p[n]$), or unvoiced ($e[n]$)) and the filter $h[n]$

$$x[n] = u[n] * h[n], \;\; 0 \leq n \leq L-1, \tag{3.10}$$

where h[n] is the vocal tract response.

Here the significance of cepstral analysis becomes apparent since eq. (3.10) can be converted into a sum via a homomorphic transformation

$$\hat{c}[n] = D_*(x[n]) = \hat{u}[n] + \hat{h}[n].$$

Cepstral analysis was first applied to speech by Michael Noll. In 1964 in published two papers in the *Journal of Acoustical Society of America* ([18, 17]) in which he applied cepstral processing techniques to short-time segments of speech signals to successfully perform pitch detection [19]. Nolls pitch detection algorithm computes a sequence of short-time cepstra and searches each cepstrum for a peak in the region of the expected pitch period. A strong peak suggests voiced speech and the location gives an estimate of the pitch period. In this way Noll was able to distinguish between voiced and unvoiced speech [2].

This success suggested to Oppenheim, Schafer, and Stockham that homomorphic deconvolution could be used to deconvolve speech. That is, by applying the inverse characteristic system for convolution to the liftered cepstrum, they would be able to separate the periodic glottal pulse excitation and the vocal tract impulse response [19].

This application of homomorphic deconvolution led to the development of the homomorphic vocoder, an analysis/synthesis speech coding system. Considering short-time segments of speech, this system combines a cepstral pitch detector with a homomorphic deconvolution to estimate the pitch period and the time varying vocal tract impulse response. These parameters can then be used to synthesize the speech signal, using the speech model shown in Figure 3.11.

In this project we are looking for connections between an *attended* speech signal and an EEG signal. As such, it is unnecessary for us to find the vocal tract impulse response explicitly. Instead, we are able to make comparisons between the low-quefrency regions of the cepstrums of the respective signals directly and so do not need to apply the inverse characteristic system for convolution (Figure 3.3). In the following chapters we describe in more detail the way we used these cepstral processing techniques to address our question, which involves both speech and EEG signals. In chapter 5 we describe a simulation of the cocktail party problem and our first approach to applying cepstral processing techniques to find connections between an attended speech signal and an EEG signal. In chapter 6 we describe the techniques used when handling real data.

Chapter 4

Stimulus-Reconstruction

While listening to speech, the cortical activity of a subject changes in response to the speech signal. Since the response occurs after the stimulus it is clear that there will be some kind of delay between the time that the speech stimulus and the neural response occurs. The stimulus-reconstruction approach addresses this and attempts to reconstruct an estimate of an input stimulus, s, using a response, R, through a linear reconstruction model, g [21]. This is a type of LTI system [8] (see section 2.3), and although the human brain is not a linear system, certain assumptions can be made which allow it to be modelled as one [4, 5].

In early studies investigating neural responses to auditory stimuli focused on brief, isolated stimuli such as individual phonemes or syllables. Stimulus-reconstruction was developed as a more direct way to investigate neural entrainment to continuous stimuli and has been used in a number of studies to model speech processing with intercranial and non-invasive electrophysiology [15, 22, 8, 10, 9]. Stimulus-reconstruction has also been used specifically to model and predict selective attention in a multispeaker environment [21].

The stimulus-reconstruction model considers a linear mapping, g, from the neural response, r, to the speech stimulus, s [8, 21]. Since g is simply a linear mapping from the *stimulus* to the *response*, these can be chosen arbitrarily. Previous applications of the stimulus-reconstruction method have used the speech envelope as the stimulus and the EEG signal as the response. However, of interest here is the connection between the cepstral coefficients of the attended speech and those of the EEG. As such, we take the cepstral coefficients of the attended speech signal as the stimulus, and the cepstral coefficients of the EEG signal as the response.

This is done by breaking the speech and EEG signals into non-overlapping time frames and calculating cepstral coefficients for each of these frames. Having calculated the cepstrum for a given sequence, the first coefficient m_0 is discarded as it corresponds to an impulse.

From the remaining cepstral coefficients, let M be the number of cepstral coefficients that were kept from each sequence. The frames were then considered as time points, with each frame giving M cepstral coefficients.

In our model, we take the stimulus, s, to be the cepstral coefficients of the attended speech signal. Therefore we let $\mathbf{s}(k)$ indicate the set of M cepstral coefficients for the k-th frame of the speech signal, and write

$$\mathbf{s}(k) = \begin{bmatrix} s(k,1) \\ \vdots \\ s(k,m) \\ \vdots \\ s(k,M) \end{bmatrix}. \tag{4.1}$$

The cepstral coefficients of the EEG signal corresponding to the speech signal are considered as the response. Here we need to take into account the fact that there are N channels. Thus we break the EEG signal into non-overlapping time frames, as we did for the speech signal, but in this case we do it for each respective channel. For each frame of each channel, we calculate the cepstral coefficients, discard the very first, m_0, and save the next M coefficients. Therefore, we let $r(k,m,n)$ indicate the m-th cepstral coefficient of the k-th frame of the n-th EEG channel, and use $\mathbf{r}(k)$ (note that this is bold face) to indicate the collection of all M cepstral coefficients for all N channels for frame k. That is, we have

$$\mathbf{r}(k) = \begin{bmatrix} r(k,1,1) & r(k,1,2) & \cdots & r(k,1,n) & \cdots & r(k,1,N) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ r(k,m,1) & r(k,m,2) & \cdots & r(k,m,n) & \cdots & r(k,m,N) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ r(k,M,1) & r(k,M,2) & \cdots & r(k,M,n) & \cdots & r(k,M,N) \end{bmatrix}. \tag{4.2}$$

We then consider the reconstruction model, g(τ,n), which represents a linear mapping from the neural response, $r(t,m,n)$, to the speech stimulus, $s(k,m)$. This is written as:

$$\hat{s}(k,m) = \sum_{\tau}\sum_{n} r(k-\tau,m,n)\cdot g(\tau,n) = \mathbf{Rg}, \tag{4.3}$$

where $\hat{s}$ is the reconstructed stimulus. $\mathbf{R}$ is the lag matrix which is written as

$$\mathbf{R} = \begin{bmatrix} \mathbf{r}(1-\tau_{\min}) & \mathbf{r}(-\tau_{\min}) & \cdots & \mathbf{r}(1) \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{r}(K) & \vdots & \cdots & \vdots \\ 0 & \mathbf{r}(K) & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \mathbf{r}(K) \end{bmatrix}, \tag{4.4}$$

where K is the total number of frames. Thus, in matrix form, we write

$$\begin{bmatrix} \mathbf{r}(1-\tau_{\min}) & \mathbf{r}(-\tau_{\min}) & \cdots & \mathbf{r}(1) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{r}(K) & \vdots & \cdots & \vdots \\ 0 & \mathbf{r}(K) & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \mathbf{r}(K) \end{bmatrix} \cdot \begin{bmatrix} \mathbf{g}(\tau_{min}) \\ \mathbf{g}(\tau_{min}+1) \\ \vdots \\ \mathbf{g}(\tau_{max}) \end{bmatrix} = \begin{bmatrix} \mathbf{s}(1) \\ \mathbf{s}(2) \\ \vdots \\ \mathbf{s}(K) \end{bmatrix},$$

where,

$$\mathbf{g}(\tau) = \begin{bmatrix} g(\tau,1) \\ \vdots \\ g(\tau,n) \\ \vdots \\ g(\tau,N) \end{bmatrix}.$$

By including the lags, the model takes into account the fact that the response to the stimulus may occur after some delay. Since we have broken the signals into frames, we consider the frames as our time points. As a result, our lags represent lags in frames. Negative lags correspond to the number of frames after the given time point. So a lag of $\tau = -2$ would represent a frame in the response two time frames after the time frame of the considered stimulus frame. We assume that the neural signal occurs as as response to the speech and not the other was round and therefore only consider negative lags in this model. The range over which we chose to consider the lags was from 0 to 450ms based on the range typically used to capture the cortical response. [8] This was calculated based on the length of frame being considered. For example, if the frame length is $25ms$, our range of lags was $\{-19:0\}$. In this example, a lag of $\tau = -2$ would represent a frame in the EEG response, 50ms after the speech stimulus being considered.

The function g is estimated by minimising the MSE between $s(k,m)$ and $\hat{s}(k,m)$:

$$\min\ \mathbf{e} = \min \sum_{k}\sum_{m} \left[s(k,m) - \hat{s}(k,m)\right]^2. \tag{4.5}$$

This is found using the following equation:

$$\mathbf{g} = \left(\mathbf{R}^{\mathrm{T}}\mathbf{R}\right)^{-1}\mathbf{R}^{\mathrm{T}}\mathbf{s} \tag{4.6}$$

where $\mathbf{R}$ is the lagged time series of the response matrix, $\mathbf{r}$.

4.1 Classification

Now we have the reconstruction model, $\mathbf{g}$, we can take an EEG signal and make a prediction, $\hat{\mathbf{s}}$, of the cepstral coefficients of the speech signal, $x[n]$, that led to that neural response. In this way we can obtain our prediction $\hat{\mathbf{s}}$ using eq. (4.3).

Assume that we have two speech signals, $x_1[n]$ and $x_2[n]$ of which we will only attend one of them. This will be our ground truth that will be defined as

$$attention_{truth} = \begin{cases} 1, & \text{if subject is attending to speech stream 1} \\ 2, & \text{if subject is attending to speech stream 2.} \end{cases} \tag{4.7}$$

In this way, the attending speech stream will be referred as *attended* and the other one as *unattended*. Let $\mathbf{s}_A$ and $\mathbf{s}_U$ denote the set of cepstral coefficients for the *attended* and *unattended* speech signal respectively, obtained as in eq. (4.1) (*stimulus*). Assume that we also have an EEG signal from a subject listening to both speech signals, $x_1[n]$ and $x_2[n]$ simultaneously. Then we can obtain the set of cepstral coefficients for these EEG data (*response*). The reconstruction model, $\mathbf{g}$ can be accomplished by means of eq. (4.6). Using this transference function, $\mathbf{g}$ and the EEG data a prediction $\hat{\mathbf{s}}$ can be made using eq. (4.3).

Prediction $\hat{\mathbf{s}}$ can be compared with the two different speech streams $\mathbf{s}_A$ and $\mathbf{s}_U$ respectively. This comparison can be made by computing the normalised mean square error (NMSE) and Pearson correlation coefficient (ρ).

The NMSE is defined as

$$\text{NMSE} = 1 - \frac{|s - \hat{s}|^2}{|s - \bar{s}|} \tag{4.8}$$

where

$$\bar{s} = \frac{1}{N}\sum_i s_i, \tag{4.9}$$

and $\hat{s}$ is the approximation of s. It therefore provides a measure of how close the prediction, $\hat{s}$ (eq. (4.3)), is from the *attended* or *unattended* stimulus respectively. The NMSE takes values in the interval $(-\infty, 1]$, the value 1 being a perfect match and 0 meaning there is no difference between the fitted sequence and a straight line [14]. In order to measure the linear correlation between the prediction $\hat{s}$ and the *attended* or *unattended* stimulus, the Pearson correlation coefficient ρ was used.

We will obtain two values for the NMSE. The first one denoted as $NMSE_{attended}$ for the *attended* speech stream, $\mathbf{s}_A$, compared to the prediction, $\hat{\mathbf{s}}$. The second one denoted as $NMSE_{unattended}$ for the *unattended* speech stream, $\mathbf{s}_U$ compared to the prediction, $\hat{\mathbf{s}}$

$$NMSE_{attended} = NMSE(\mathbf{s}_A, \hat{\mathbf{s}}) \tag{4.10}$$

$$NMSE_{unattended} = NMSE(\mathbf{s}_U, \hat{\mathbf{s}}). \tag{4.11}$$

In a similar way, we will obtain two values for the Pearson ρ. The first one denoted as $\rho_{attended}$ for the *attended* speech stream, $\mathbf{s}_A$, compared to the prediction, $\hat{\mathbf{s}}$. The second one denoted as $\rho_{unattended}$ for the *unattended* speech stream, $\mathbf{s}_A$ compared to the prediction, $\hat{\mathbf{s}}$

$$\rho_{attended} = corr(\mathbf{s}_A, \hat{\mathbf{s}}) \tag{4.12}$$

$$\rho_{unattended} = corr(\mathbf{s}_U, \hat{\mathbf{s}}). \tag{4.13}$$

The NMSE and ρ values give us a prediction (indication) of which speech stream the listener was attending to. The speech stream with the highest NMSE and ρ values is deemed to be the $\widehat{attention}$ speech stream,

$$\widehat{attention}_{NMSE} = \begin{cases} 1, & \text{if } NMSE_{attended} > NMSE_{unattended} \\ 2, & \text{if } NMSE_{attended} < NMSE_{unattended} \end{cases} \tag{4.14}$$

$$\widehat{attention}_{\rho} = \begin{cases} 1, & \text{if } \rho_{attended} > \rho_{unattended} \\ 2, & \text{if } \rho_{attended} < \rho_{unattended} \end{cases} \tag{4.15}$$

Each of these indications of $\widehat{attention}$ speech stream is then compared with the true result of which speech stream was actually being attended ($attention_{truth}$). This comparison will give a classification outcome *correct* (=1) or *incorrect* (=0),

$$classification = \begin{cases} 1, & \text{if } attention_{truth} = \widehat{attention} \\ 0, & \text{otherwise .} \end{cases} \tag{4.16}$$

For this project, we applied this stimulus-reconstruction using cepstral coefficients method to real data. These data contained information from different subjects. Each subject with 30 trials that consisted of two speech streams and EEG recordings for 128 channels. Subjects were asked to attend one of the two recordings that were played simultaneously to their right and left ears respectively. A classification was possible for each trial of each subject. The classification rates were then based on the number of these 30 results that were correct. These results will be discussed in further detail in chapter 6.

Chapter 5

Simulation of the Cocktail Party Problem

In this chapter we outline our preliminary investigation into the connection between the cepstral coefficients of an EEG signal and those of an *attended* speech signal. We began our analysis by considering the fundamentals of the problem being considered. Our aim was to use cepstral processing to identify listener attention in a scenario where multiple speakers are competing at the same time. We began by creating simulations of the problem with varying levels of complexity in order to get a better understanding of how cepstral analysis could be applied to this problem and to test the limitations of different spectral estimation methods.

In our first simulations we simulated speech, which will be referred to as the *attended speech*, based on the idea of speech being a convolution of a glottal impulse train and a vocal tract response (see section 3.4). Based on the assumption that the EEG response follows the speech stimulus, we simulated EEG by adding noise to the simulated speech signal. This simulated EEG signal will be referred to as EEG_{sim}. This was done a number of times with increasing amounts of added noise. Two types of noise were considered separately, *white noise*, and *EEG noise* which we simulated using a resting EEG recording. Then, calculating the cepstral coefficients of the simulated speech and EEG signals over short-time intervals (*frames*), we compared them to see how closely they matched. Clearly, with no added noise the simulated EEG is simply the speech signal, and as such we would obtain a perfect match between the cepstral coefficients of the *attended speech* and the EEG_{sim}. We then proceeded to add more noise to find the point at which the cepstral coefficients no longer matched well, and to compare different spectral estimation methods used in the calculation of the cepstrum.

To do this, the cepstral coefficients of the *attended speech*, and the EEG_{sim}, were obtained and compared using normalised mean square error (NMSE) and Pearson correlation coefficient (ρ). The NMSE provides a measure of how close the cepstral coefficients from the *attended speech*$_{sim}$ were to the cepstral coefficients of EEG_{sim}. In order to measure the linear correlation between the cepstral coefficients of the two signals, the Pearson correlation coefficient ρ was used. These NMSE and ρ values (defined in section 4.1) were found for different levels of added noise in EEG_{sim} to give an idea of the noise level at which the cepstral coefficients of the EEG_{sim} no longer matched those of the *attended speech*.

Two different types of *attended speech* were used. For the first stage of the simulations, a simulated signal, which will be denoted as *attended speech*$_{sim}$, was obtained using an impulse train and an impulse response. For the second stage, a recording of a female voice saying the word in spanish "Hola" was used; it will be denoted as *attended speech*$_{real}$.

Two types of EEG_{sim} were considered. The first was obtained by adding a realization of *white noise* to each *attended speech*, we will refer to this sequence as EEG_{sim_1}. The second, EEG_{sim_2}, was obtained by adding a realization of *simulated background EEG* to each *attended speech*. In the first stage of this chapter we used a single 20*ms* frame for each of the signals. Later on, we extended the analysis and considered all the 20*ms* frames in the whole length of the signals.

These simulation experiments were a preliminary investigation into how we would approach the problem of determining listening attention using cepstral analysis. Although they are basic, and perhaps very simplistic in comparison to the real data, these simulations helped us to gain insight into how we would build our model and allowed us to start writing Matlab scripts in more easily handled stages. They also suggested the best frame lengths and spectral estimation methods to use when modelling using real data.

5.1 Simulated Speech and Simulated EEG - Single 20ms Frame

In chapter 3 we discussed in more detail the reasoning of why speech signals can be modelled as an impulse train and impulse response. With this idea in mind for the first experiment, we considered simulated speech and EEG signals with different types and levels of noise. The simulated speech signals were created using the impulse response of a linear filter $h[n]$, which was convolved with an impulse train $e[n]$ (see Figure 5.1). *attended speech*$_{sim}$ was simulated with 10,000 data points. Since for this exercise we are simulating signals that are 1-second long, it follows that the sampling frequency for *attended speech*$_{sim}$ is 10,000Hz.

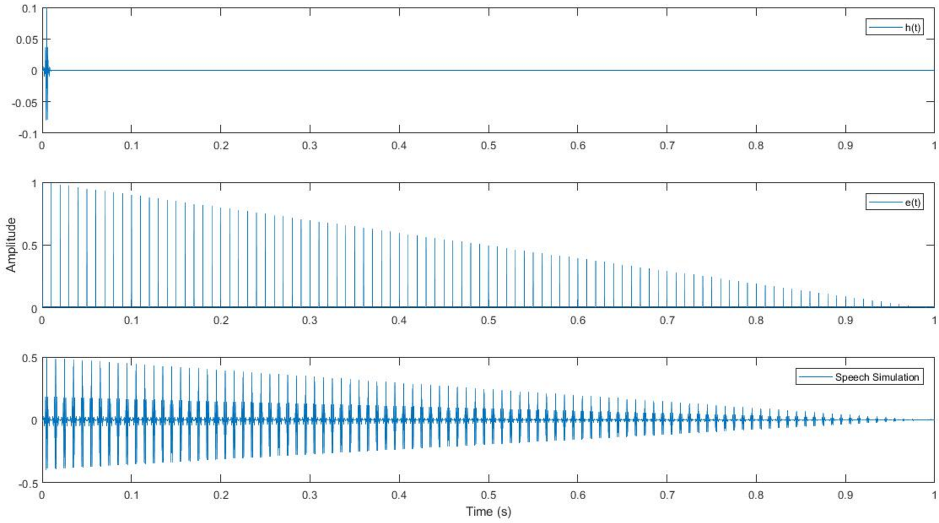


Figure 5.1 Simulation of 1 second of $attended\ speech_{sim}$ obtained from the convolution of an impulse response $h(t)$ with the impulse train $e(t)$

Two types of simulated noise were considered: *white noise*, and *simulated background EEG*. The *white noise* sequence was obtained by generating 1,000 random numbers from a $\mathbf{N}(0,1)$ distribution. The *simulated background EEG* was obtained from fitting an $AR(15)$ model to EEG recordings taken from a single channel close to the left ear. Figure 5.2(a) shows a simulation of *white noise* and Figure 5.2(b) shows the simulated sequence $attended\ speech_{sim}$. For this exercise we simulated signals that were considered to be 1-second long. The 1,000 *noise* points correspond to a sampling frequency of 1000Hz, in order to match this frequency with the one from the $attended\ speech_{sim}$ (10kHz), up-sampling using interpolation was performed on the noise signals by re-sampling the noise sequence at 10 times the original sampling rate. Once the two signals had the same number of points and the same sampling frequency, the two noise signals were added to the attended speech signal respectively giving

$$EEG_{sim_1} = attended\ speech_{sim} + white\ noise$$
$$EEG_{sim_2} = attended\ speech_{sim} + simulated\ background\ EEG.$$

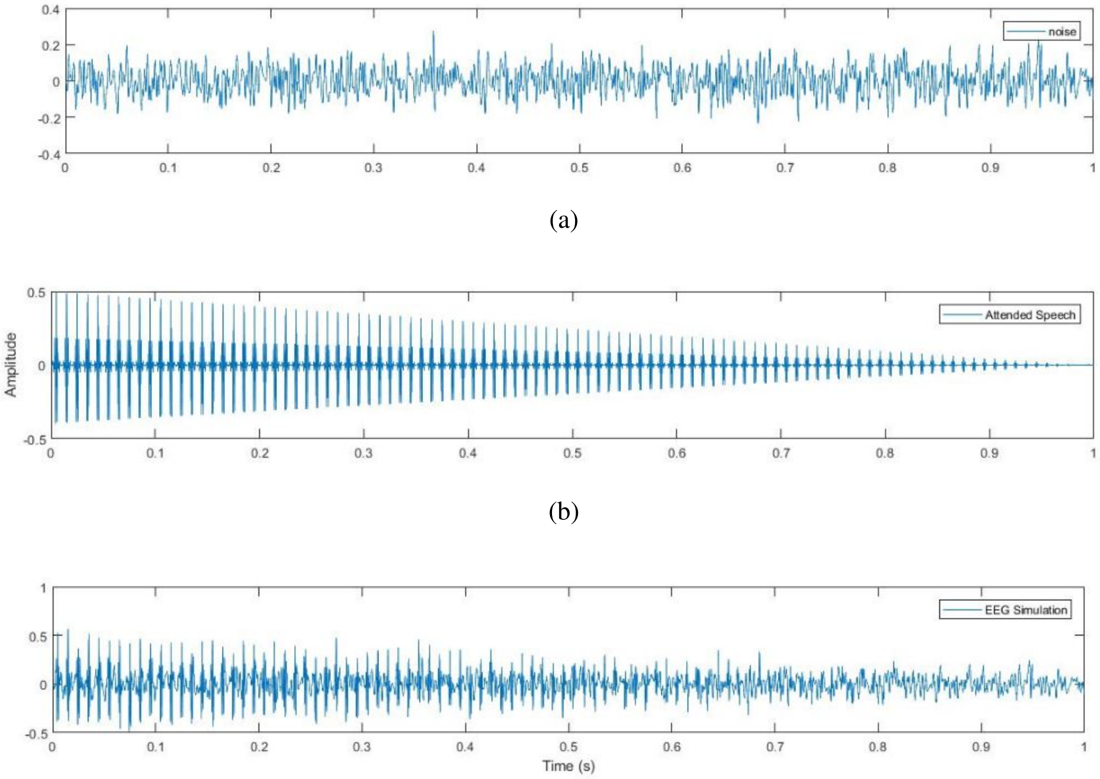
The two types of noise were scaled up to different levels of signal-to-noise ratio (SNR) to identify up to which point it was still possible to identify the cepstrum coefficients as described above. For a clearer measure of the different levels of *noise*, a scaling factor σ was derived from the Signal-to-Noise Ratio

$$SNR = 10\log_{10}\frac{P_{signal}}{\sigma * P_{Noise}}, \tag{5.1}$$

that is,

$$\sigma = \frac{P_{signal}}{P_{Noise} * 10^{\frac{SNR}{10}}}. \tag{5.2}$$

The σ corresponding for different levels of SNR was obtained. Figure 5.2 (c) includes a visualization of the *noise* scaled by the factor that corresponds to an SNR of -10.



(a)

(b)

(c)

Figure 5.2 Figure (a) shows a realization of the up-sampled *noise*, Figure (b) is the *attended speech*$_{sim}$, and Figure (c) is the *EEG*$_{sim}$ obtained by adding together these two signals from (a) and (b) using SNR = -10.

Following the procedure detailed in section 3.1, a 20*ms* frame was taken from the signals, and the cepstral coefficients were obtained for the *attended speech*$_{sim}$ and *EEG*$_{sim}$ using the following PSD estimation methods:

- FFT using a Gaussian window

- Periodogram
- Periodogram with a Hanning window
- Welch method
- Multitaper

Having calculated the cepstrum for a given sequence, the first coefficient m_0 is discarded as it corresponds to an impulse. From the remaining cepstral coefficients, let M be the number of cepstral coefficients that were kept from each sequence. For this exercise we used $M = 13$, i.e. the first 13 cepstral coefficients of each sequence (after coefficient m_0) from these $20ms$ frames were kept and then compared with each other using the NMSE and Pearson ρ value as described above. The following tables show the mean of the NMSEs and the ρ for the 10,000 realizations.

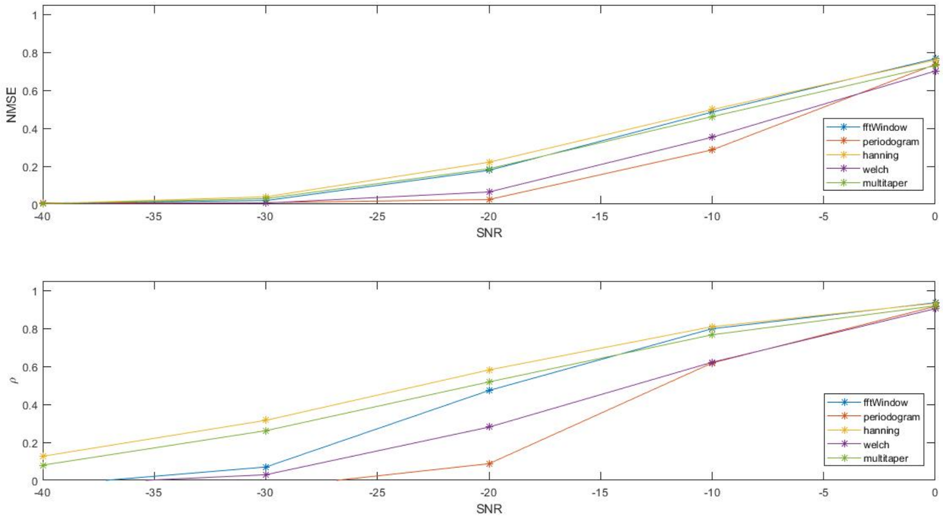


Figure 5.3 Mean of NMSE and ρ of the fit from the *attended speech*$_{sim}$ cepstral coefficients and the EEG_{sim_1} (*white noise*) cepstral coefficients, using different PSDE methods. The $x-axis$ shows the SNR ranging from -40 to 0.

Table 5.1 Mean of NMSE for 10,000 realizations of *attended speech*$_{sim}$ and EEG_{sim_1}

PSDE	SNR = -40	SNR = -30	SNR = -20	SNR = -10	SNR = 0
FFT	0.0085	0.0199	0.1595	0.4584	0.7452
Periodogram	0.0257	0.0266	0.0738	0.3624	0.7284
Hanning	0.0008	0.0258	0.2006	0.4804	0.7516
Welch	0.0059	0.0128	0.1149	0.4239	0.7625
Multitaper	0.0016	0.0450	0.2475	0.5218	0.7665

Table 5.2 Mean of ρ for 10,000 realizations of *attended speech*$_{sim}$ and EEG_{sim_1}

PSDE	SNR = -40	SNR = -30	SNR = -20	SNR = -10	SNR = 0
FFT	- 0.0155	0.0757	0.4492	0.7879	0.9333
Periodogram	- 0.0141	- 0.0079	0.1930	0.7289	0.9323
Hanning	0.1336	0.3156	0.5764	0.8113	0.9370
Welch	0.0008	0.0875	0.3673	0.6999	0.9413
Multitaper	0.0697	0.3099	0.6037	0.8226	0.9435

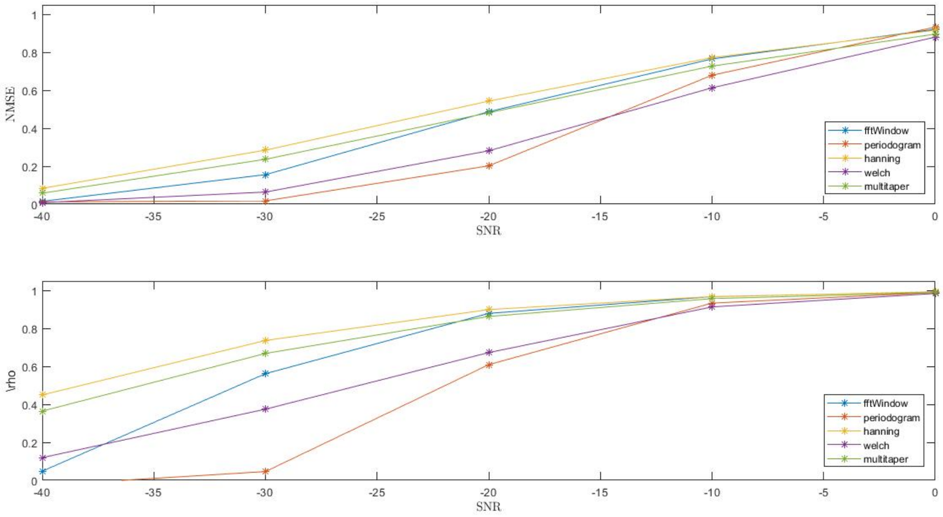


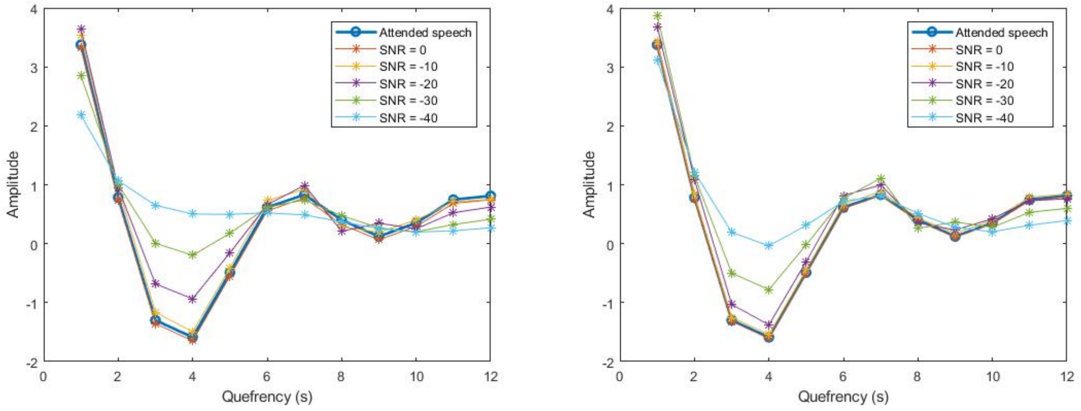
Figure 5.4 Mean of NMSE and ρ of the fit from the *attended speech*$_{sim}$ cepstral coefficients and the EEG_{sim_2} (*simulated background EEG*) cepstral coefficients, using different PSDE methods. The $x-axis$ shows SNR from -40 to 0.

Table 5.3 Mean of NMSE for 10,000 realizations of *attended speech*$_{sim}$ and EEG_{sim_2}

PSDE	SNR = -40	SNR = -30	SNR = -20	SNR = -10	SNR = 0
FFT	0.0218	0.1828	0.5007	0.7689	0.9252
Periodogram	0.0218	0.0658	0.3511	0.7280	0.9193
Hanning	0.0760	0.2735	0.5375	0.7760	0.9251
Welch	0.0178	0.1120	0.3705	0.7058	0.9335
Multitaper	0.0848	0.3032	0.5540	0.7768	0.9189

Table 5.4 Mean of ρ for 10,000 realizations of *attended speech*$_{sim}$ and EEG_{sim_2}

PSDE	SNR = -40	SNR = -30	SNR = -20	SNR = -10	SNR = 0
FFT	0.1299	0.6188	0.8967	0.9761	0.9962
Periodogram	0.0197	0.2477	0.8038	0.9642	0.9940
Hanning	0.4853	0.7597	0.9138	0.9771	0.9961
Welch	0.2015	0.4660	0.7626	0.9528	0.9938
Multitaper	0.4552	0.7670	0.9125	0.9764	0.9960



(a) (b)

Figure 5.5 First 13 cepstral coefficients of a realization from the sequences: *attended speech*$_{sim}$ and EEG_{sim_1}. Figure (a) corresponds to one realization using *white noise* and figure (b) corresponds to one realization using *simulated background EEG*. From these graphs we can see how the coefficients from the *attended speech*$_{sim}$ closely match the coefficients from EEG_{sim_1} and EEG_{sim_2}

Looking at the NSMEs from Tables 5.1 and 5.3, we can say that among the selected PSD estimation methods, the periodogram with a Hanning window gives the best fit for the coefficients of the two frames. This is also confirmed by the ρ values shown in Tables 5.2 and 5.4. It can also be seen from Figure 5.5 that once the SNR levels are between -30 and -40, it becomes complicated to assess the goodness of fit.

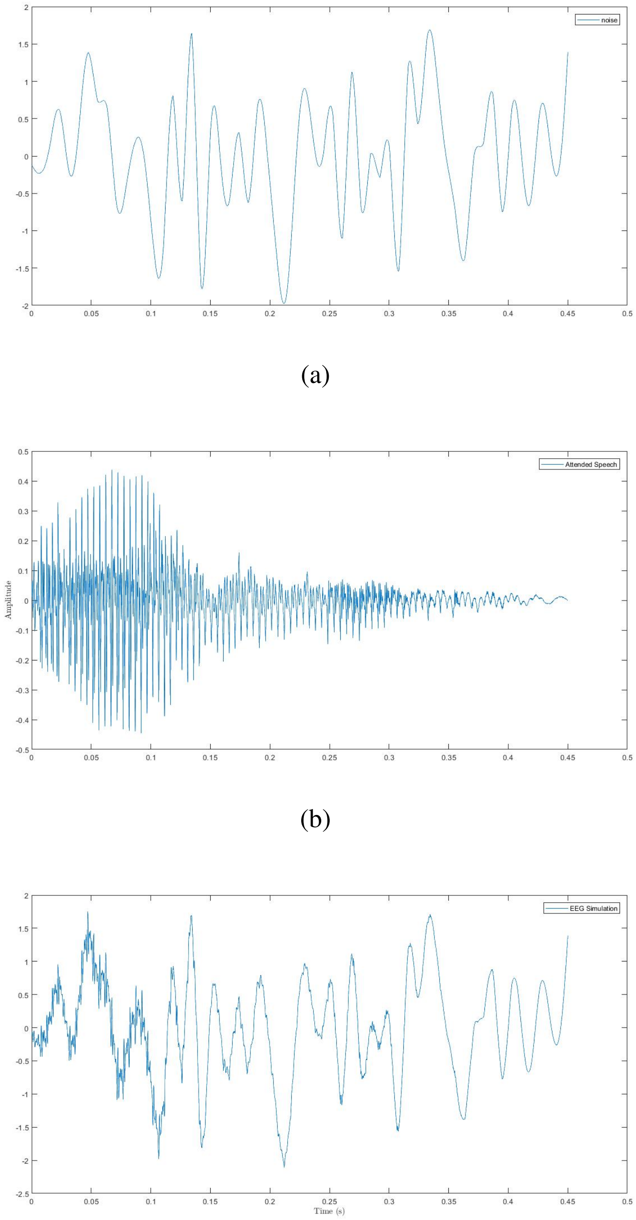
5.2 Real Speech and Simulated EEG - single 20ms frame

For this second set-up, a real speech sequence was considered instead of simulated speech. We will refer to to this real speech sequence as *attended speech*$_{real}$. Two different types of *noise* were also simulated and added to the speech stream at various SNR levels in order to obtain simulated EEG. These *noise* simulations were generated assuming a sampling frequency of 10,000Hz. For the *attended speech*$_{real}$ we considered a recording of a female voice saying the Spanish word "Hola". Two types of *noise* were added to the real speech signal to obtain two EEG simulations

$$EEG_{sim_1} = attended\ speech_{real} + white\ noise \tag{5.3}$$

$$EEG_{sim_2} = attended\ speech_{real} + simulated\ background\ EEG. \tag{5.4}$$

Figure 5.6(a) shows a simulation of *white noise* and Figure 5.6(b) shows the simulated sequence *attended speech*$_{real}$. Figure 5.6 (c) includes a visualization of the *noise* scaled by the factor that corresponds to an SNR of -20.



(a)

(b)

(c)

Figure 5.6 Figure (a) shows a realization of the *noise*, Figure (b) shows the *attended speech*$_{sim}$, and Figure (c) is the *EEG*$_{sim}$ obtained by adding together the two signals from (a) and (b) using SNR = -20.

From these signals a frame of 20*ms* length was extracted. The same procedure detailed in section 3.1 was followed, for each frame the PSD was estimated using the following methods:

- FFT using a Gaussian window
- Periodogram
- Periodogram with a Hanning window
- Welch method
- Multitaper

In a similar way as done in the previous section, the first 13 cepstral coefficients for each sequence were kept and then compared with each other using the NMSE and correlation ρ.

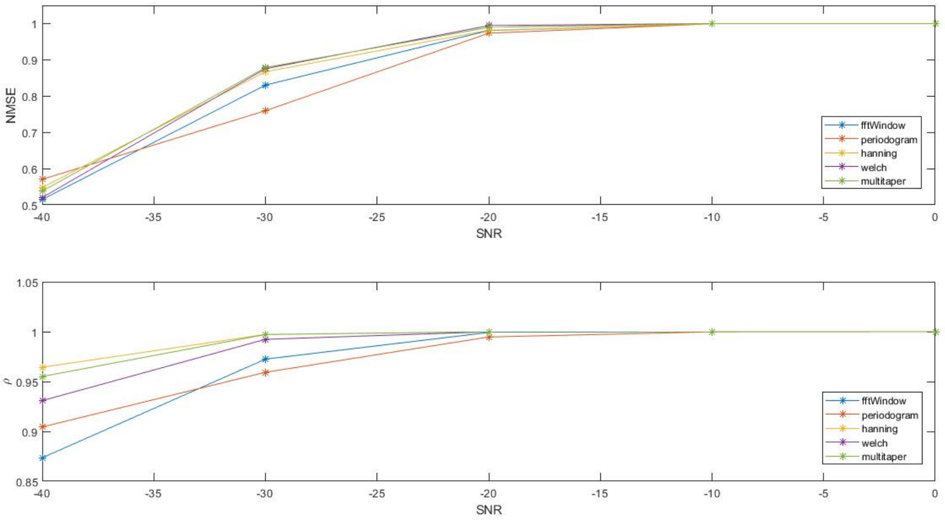


Figure 5.7 Mean of NMSE and ρ of the fit from the *attended speech*$_{real}$ cepstral coefficients and the EEG_{sim_1} (*white noise*) cepstral coefficients, using different PSDE methods. The $x-axis$ shows SNR from -40 to 0.

Table 5.5 Mean of NMSE for *attended speech*$_{real}$ and 10,000 realizations of EEG_{sim_1}

PSDE	SNR = -40	SNR = -30	SNR = -20	SNR = -10	SNR = 0
FFT	0.5149	0.8303	0.9808	0.9993	1.0000
Periodogram	0.5704	0.7598	0.9733	0.9993	1.0000
Hanning	0.5476	0.8672	0.9816	0.9993	1.0000
Welch	0.5204	0.8755	0.9943	1.0000	1.0000
Multitaper	0.5379	0.8787	0.9895	0.9999	1.0000

Table 5.6 Mean of ρ for *attended speech*$_{real}$ and 10,000 realizations of EEG_{sim_1}

PSDE	SNR = -40	SNR = -30	SNR = -20	SNR = -10	SNR = 0
FFT	0.8736	0.9725	0.9992	1.0000	1.0000
Periodogram	0.9045	0.9593	0.9947	0.9998	1.0000
Hanning	0.9644	0.9973	0.9999	1.0000	1.0000
Welch	0.9309	0.9923	0.9999	1.0000	1.0000
Multitaper	0.9551	0.9972	1.0000	1.0000	1.0000

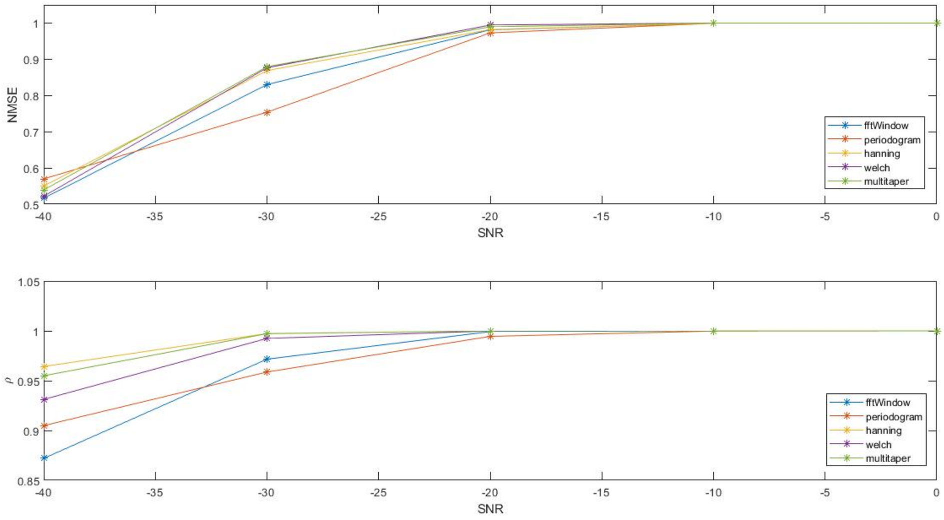


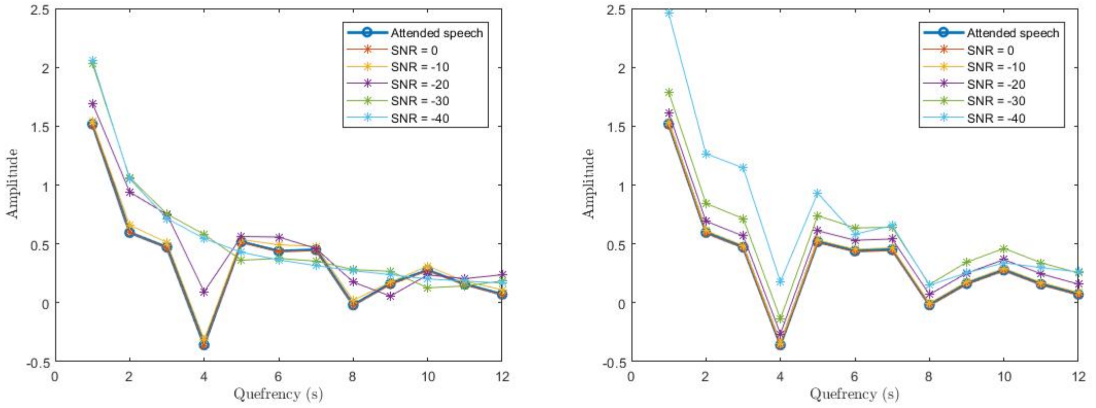
Figure 5.8 Mean of NMSE and ρ of the fit from the *attended speech*$_{real}$ cepstral coefficients and the *EEG*$_{sim_2}$ (*white noise*) cepstral coefficients, using different PSDE methods. The $x-axis$ shows SNR from -40 to 0.

Table 5.7 Mean of NMSE for *attended speech*$_{real}$ and 10,000 realizations of *EEG*$_{sim_2}$

PSDE	SNR = -40	SNR = -30	SNR = -20	SNR = -10	SNR = 0
FFT	0.5165	0.8301	0.9813	0.9992	1.0000
Periodogram	0.5699	0.7540	0.9728	0.9993	1.0000
Hanning	0.5503	0.8687	0.9821	0.9991	1.0000
Welch	0.5221	0.8764	0.9944	1.0000	1.0000
Multitaper	0.5393	0.8795	0.9896	0.9999	1.0000

Table 5.8 Mean of ρ for *attended speech*$_{real}$ and 10,000 realizations of *EEG*$_{sim_2}$

PSDE	SNR = -40	SNR = -30	SNR = -20	SNR = -10	SNR = 0
FFT	0.8720	0.9716	0.9992	1.0000	1.0000
Periodogram	0.9049	0.9587	0.9945	0.9998	1.0000
Hanning	0.9642	0.9973	0.9999	1.0000	1.0000
Welch	0.9312	0.9924	0.9999	1.0000	1.0000
Multitaper	0.9550	0.9972	1.0000	1.0000	1.0000



(a)

(b)

Figure 5.9 First 13 cepstral coefficients of the sequences: *attended speech*$_{sim}$ and EEG_{sim_1}. Figure (a) corresponds to one realization using *white noise* and figure (b) to one realization using *simulated background EEG*.

We can see from Tables 5.5 to 5.8, that when using real speech both the NMSE and the ρ values were higher compared to those obtained using simulated speech streams. Even for SNR levels of -30 (for both *white noise* and *simulated background EEG*) the NMSE values were above 0.8 (except for the periodogram). For all the SNR levels considered in this exercise, the obtained ρ values were above 0.85.

5.3 Real Speeches and Simulated EEG - whole sequence

In order to have a similar scenario to the one described as the cocktail party problem two real speech streams were considered for this section. One speech stream was the recording of a female voice saying the Spanish word "Hola" used in the previous section, we will refer to this stream as the *attended speech*. A second speech stream consisted of the recording of a male voice saying the Spanish word "Hola", we will refer to this stream as the *unattended speech* (see Figure 5.10). We added simulated *noise* to the *attended speech* to obtain a simulated EEG recording, EEG_{sim}. In this way we obtained the next 3 sequences:

1. *attended speech*
2. *unattended speech*
3. EEG_{sim} = *attended speech* + *noise*

We computed the cepstral coefficients from a frame length $20ms$ of the sequence EEG_{sim} and compared them with the cepstral coefficients from frame length $20ms$ of the sequence *attended speech* and *unattended speech* respectively.

The idea behind this exercise was to see how well for different levels of SNR the cepstral coefficients from EEG_{sim} would fit the cepstral coefficients of the *attended speech* rather than those of the *unattended speech*.

Similarly, as in the previous section, two types of noise were considered:

$$EEG_{sim_1} = attended\ speech + white\ noise \tag{5.5}$$

$$EEG_{sim_2} = attended\ speech + simulated\ background\ EEG \tag{5.6}$$

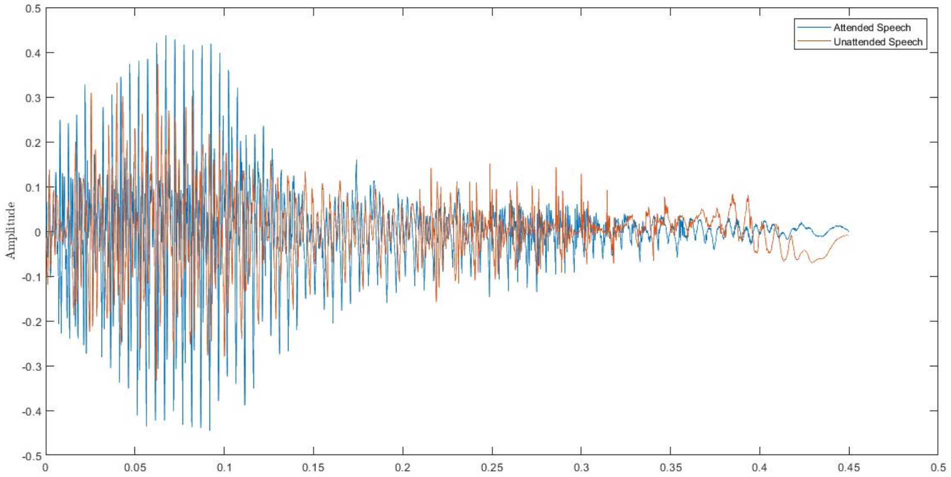


Figure 5.10 Recordings of voices saying the Spanish word "Hola". The line in blue corresponds to the *attended speech* and the line in red to the *unattended speech*.

The speech and EEG signals were broken into non-overlapping frames and the cepstral coefficients were calculated for each of these frames. At this stage in the investigation we were not considering lags between the speech *stimulus* and the EEG *response*. However, we use part of the method outlined in chapter 4. Using eqs. (4.1) and (4.2) we let $\mathbf{s}_A(k)$, $\mathbf{s}_U(k)$, and $\mathbf{r}(k)$ indicate the set of M cepstral coefficients for the k-th frame of the *attended*, *unattended speech*, and EEG_{sim} signals respectively. That is

$$\mathbf{s}_A(k) = \begin{bmatrix} s_A(k,1) \\ \vdots \\ s_A(k,m) \\ \vdots \\ s_A(k,M) \end{bmatrix}, \tag{5.7}$$

$$\mathbf{s}_U(k) = \begin{bmatrix} s_U(k,1) \\ \vdots \\ s_U(k,m) \\ \vdots \\ s_(k,M) \end{bmatrix}, \tag{5.8}$$

and

$$\mathbf{r}(k) = \begin{bmatrix} r(k,1,1) \\ \vdots \\ r(k,m,1) \\ \vdots \\ r(k,M,1) \end{bmatrix}. \tag{5.9}$$

Note that since we are not simulating a multi-channel EEG signal, there is only one EEG channel, thus N = 1. Each of these sets of $M = 13$ cepstral coefficients were then stacked into column vectors of length $K \times M$.

$$\mathbf{s_A} = \begin{bmatrix} s_A(1,1) \\ \vdots \\ s_A(1,m) \\ \vdots \\ s_A(1,M) \\ s_A(2,1) \\ \vdots \\ s_A(2,m) \\ \vdots \\ s_A(2,M) \\ \vdots \\ s_A(K,1) \\ \vdots \\ s_A(K,m) \\ \vdots \\ s_A(K,M) \end{bmatrix}, \mathbf{s_U} = \begin{bmatrix} s_U(1,1) \\ \vdots \\ s_U(1,m) \\ \vdots \\ s_U(1,M) \\ s_U(2,1) \\ \vdots \\ s_U(2,m) \\ \vdots \\ s_U(2,M) \\ \vdots \\ s_U(K,1) \\ \vdots \\ s_U(K,m) \\ \vdots \\ s_U(K,M) \end{bmatrix}, \text{ and } \mathbf{r} = \begin{bmatrix} r(1,1,1) \\ \vdots \\ r(1,m,1) \\ \vdots \\ r(1,M,1) \\ r(2,1,1) \\ \vdots \\ r(2,m,1) \\ \vdots \\ r(2,M,1) \\ \vdots \\ r(K,1,1) \\ \vdots \\ r(K,m,1) \\ \vdots \\ r(k,M,1) \end{bmatrix} \tag{5.10}$$

The NMSE and Pearson's ρ values were calculated between the simulated EEG and the attended and unattended speech streams respectively, for each of the cepstral coefficients of each respective frame. These NMSE and ρ values were used to determine if the sequence **r** was a better fit to sequence $\mathbf{s_A}$ or to sequence $\mathbf{s_U}$. If the NMSE and ρ between **r** and $\mathbf{s_A}$ were higher than the NMSE and ρ between **r** and $\mathbf{s_U}$, then we could conclude that EEG_{sim} was closer to the *attended speech*. This procedure was carried our with 10,000 respective realizations of EEG_{sim_1} and EEG_{sim_2} to determine how well the simulated EEG was following the *attended speech*.

Comparing the *NMSE* and ρ values obtained for each type of noise, we obtained a classification based on counting the number of times that $NMSE_{attended}$ was greater than $NMSE_{unattended}$. Similarly, we obtained a classification based on counting the number of times that $\rho_{attended}$ was greater than $\rho_{unattended}$. Using these classifications we obtained the *decoding accuracy* (%) based on each measure *NMSE* and ρ respectively and for these two types of EEG_{sim}. These *decoding accuracies* (%) are shown in the next Tables 5.9 to 5.12. From these tables we can see that a suitable PSD estimation method might be the periodogram using a Hanning window or the Welch method. For this project, we decided to use the Welch method using a Hanning window. The multitaper also performed well but it was decided not to include it because of long computation times.

Table 5.9 Decoding accuracy based on NMSE. Real speech and type of noise EEG_{sim_1}

PSDE	SNR = -40	SNR = -30	SNR = -20	SNR = -10	SNR = 0
FFT	0	0	0	100	100
Periodogram	0	0.1	3.8	100	100
Hanning	0	0	0	100	100
Welch	0	0	0	100	100
Multitaper	0	0	0	100	100

Table 5.10 Decoding accuracy based on ρ. Real speech and type of noise EEG_{sim_1}

PSDE	SNR = -40	SNR = -30	SNR = -20	SNR = -10	SNR = 0
FFT	0	0	54	100	100
Periodogram	0	0	1.5	100	100
Hanning	0	100	100	100	100
Welch	100	100	100	100	100
Multitaper	1.5	100	100	100	100

Table 5.11 Decoding accuracy based on NMSE. Real speech and type of noise EEG_{sim_2}

PSDE	SNR = -40	SNR = -30	SNR = -20	SNR = -10	SNR = 0
FFT	0	4.4	100	100	100
Periodogram	0	1.6	100	100	100
Hanning	0	0	100	100	100
Welch	0	0	100	100	100
Multitaper	0	0	100	100	100

Table 5.12 Decoding accuracy based on ρ. Real speech and type of noise EEG_{sim_2}

PSDE	SNR = -40	SNR = -30	SNR = -20	SNR = -10	SNR = 0
FFT	0	94.5	100	100	100
Periodogram	0	2.7	100	100	100
Hanning	100	100	100	100	100
Welch	23.2	100	100	100	100
Multitaper	99.9	100	100	100	100

In this chapter we presented the results of our preliminary investigation into the connection between the cepstral coefficients of an EEG signal and those of an *attended* speech signal. Various simulations of the problem with different levels of complexity were carried out in order to get a clearer idea of how cepstral analysis could be applied to this problem and to test the limitations of different spectral estimation methods. Although the simulations are simplistic in comparison to the real data, they helped us to gain an insight into the methods we would use to approach the problem and meant we could begin writing Matlab scripts in more easily handled stages. They also suggested the optimal frame lengths and spectral estimation methods to use when working with real data. We attempted to use the method described in this chapter on real data, where cepstral coefficients calculated for frames corresponding to the same time-point of the speech and EEG signals were compared. i.e. comparisons were made that do not assume any delay in the response between the speech stimulus and the EEG response. Although this appeared to work for some of the subjects, it was not consistent and did not agree with results from previous studies that show a delay between the speech stimulus and the EEG response [21, 16]. Therefore, considering the success of the stimulus-reconstruction method in the context of this cocktail-party problem, our next step was to build a stimulus-reconstruction model that used cepstral coefficients. The results of this stimulus-reconstruction model are covered in the next chapter.

Chapter 6

Results

This chapter includes a description and a summary of the obtained results when we tried stimulus-reconstruction using cepstral coefficients on real data. The utilized data were published previously by O'Sullivan et al. [21]. More details about these data can be found in appendix A. The data contain information for each subject. Each subject, (30 in total) performed 30 trials. Each trial was 1-minute and consisted in storytelling of two speech streams that were played through headphones to each subject: one was played to their left ear, and the other to their right ear. Subjects were instructed to attend to one speech stream obtaining two groups

$$attention_{truth} = \begin{cases} 1, & \text{if subject attended to speech stream on left ear} \\ 2, & \text{if subject attended to speech stream on right ear.} \end{cases} \tag{6.1}$$

Equation (6.1) was used as an indicator containing the *ground truth* for assessing predictions made by the model. In order to assess if each subject attended correctly to the indicated speech, they had to answer some questions related to the attended story [21, 23]. Results reported by O'Sullivan et al. [21] shown that $80.4 \pm 7.3\%$ of the answers related to the attended speech were correct, while $27.1 \pm 7.0\%$ of the questions related to the unattended speech were correctly answered.

Electroencephalography data was recorded for each subject using 128 electrode positions at a sampling rate of 512Hz using a BioSemi Active Two system [21]. More details about these data can be found in appendix A. Speech streams for the 30 trials were recorded using a sampling frequency of 44100 Hz. These speech signals were downsampled to 1024Hz, while EEG data was upsampled to give an equivalent sampling rate of 1024Hz. This was made in order to match both sampling frequencies of EEG data and speech signals as well to decrease computation time.

Let $x_{i,left}[n]$ and $x_{i,right}[n]$ denote the speech signal of the $i-$th of the 30 trials that was played to the left and right ear of the subjects respectively. Let $EEG_{i,j}[n]$ denote the EEG recorded for the $i-$th of the 30 trials corresponding to the $j-$th subject. As it was described in chapter 4, a set of cepstral coefficients was obtained for each speech stream $x_{i,left}[n]$ and $x_{i,right}[n]$ yielding matrices $\mathbf{s}_A$ and $\mathbf{s}_U$ (stimulus) according to each case. Cepstral coefficients were also obtained for the sequence $EEG_{i,j}[n]$ (response) and, using the lag matrix with these cepstral coefficients, matrix $\mathbf{R}$ was obtained as in eq. (4.4). The reconstructed stimulus $\hat{\mathbf{s}}$ was accomplish using eq. (4.3) and function g was estimated as in eq. (4.5). This function $\mathbf{g}$ was trained on 29 of the 30 trials and then tested on the 30th trial (for each subject independently). Taking each of the trials as a testing set respectively, and training using the remaining 29 trials, we were able to obtain 30 classification results for each subject. The classification rates were then based on the number of these 30 results that were correct.

Because of the demand in computational time and power, a first setup using certain specifications was needed. In order to assess which combination of length frame, channels, and number of cepstral coefficients might be the optimal one. By running the model with this initial setup we tried to see which combination yielded the highest classification rates using the least amount of data in the most efficient possible way. Classification rates were obtained following the process described in chapter 4. From simulations shown in chapter 5 we found that the Welch method might be a suitable first option for PSD estimation. This method was used for the first setup.

6.1 Initial Setup

Because computation times were long making it complicated to run the model for all possible combination, an initial setup needed to be chosen. The first setup that was used consisted of estimating the PSD using the Welch method for different length frames: 25, 50, and 75ms. EEG recordings from a set of 9 channels were considered: $A1$, $A19$, $A23$, $B22$, $B26$, $C17$, $C21$, $D18$, $D23$, (see fig. 6.1). Cesptral coefficients were obtained for these different lengths frames. As we did for the simulations in chapter 3, the first coefficient m_0 was discarded keeping different number of coefficients up to m_5, m_7, m_{11}, and m_{13}.

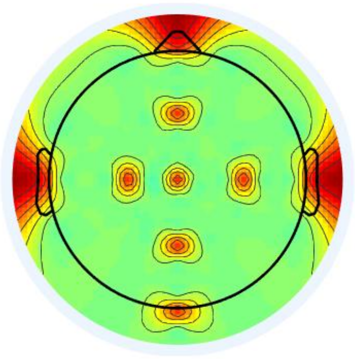


Figure 6.1 Highlighted in red channels $A1$, $A19$, $A23$, $B22$, $B26$, $C17$, $C21$, $D18$, and $D23$. These channels were selected as a first set for the initial setup.

As described at the beginning of this chapter, for each trial of each subject we first obtained the corresponding sets of cepstral coefficients, $\mathbf{s}_A$ and $\mathbf{s}_U$, as well as the corresponding lag matrix using cepstral coefficients $\mathbf{R}$. Applying stimulus-reconstruction using cepstral coefficients (eq. (4.3)) we obtained the function $\mathbf{g}$. Using this function $\mathbf{g}$ we obtained our prediction $\hat{\mathbf{s}}$. This was done for each subject and for each trial. NMSE and Pearson ρ were obtained as shown in eqs. (4.10) to (4.10). These values were then compared as in eqs. (4.14) and (4.15) to predict which speech the listener was attending to. For each subject a total of 30 classification rates were achieved based on how many of these classification were correct when compared to $attention_{truth}$ from eq. (6.1).

The following Table 6.1 shows the decoding accuracy (%) for all 30 trials of all 30 subjects based on NMSE. The maximum classification rates using NMSE were achieved when using frames with 25ms length and for all the selected values of M cepstral coefficients in this initial setup. These decoding accuracies were above 90%. For this 25ms frame length, the maximum decoding accuracy (96.34%) was achieved when $M = 13$, i.e. Having discarded the first coefficient m_0 as it corresponds to an impulse at zero, the first 13 cepstral coefficients of each frame were kept. This agrees with Huang et al. [11], they mention that in most cases the first 13 cepstral coefficients are kept when dealing with speech recognition. Table 6.1 also shows the mean of NMSE (*attended* and *unattended*) and its standard deviation (in parentheses). A Kruskal-Wallis test [12] was performed in order to determine if the obtained values for $NMSE_{attended}$ and $NMSE_{unattended}$ for all trails of all subjects with frames 25ms length were significantly different. The results of this test indicate that NMSE values for the *attended* and *unattended* speeches are significantly different (χ^2 = 45.76, p<0.001). Box plots of NMSE for all trails of all subjects using frames with 25ms length and for each value of M are shown in Figure 6.2. Box plots of NMSE for each subject's 30 trials are shown in Figure 6.3.

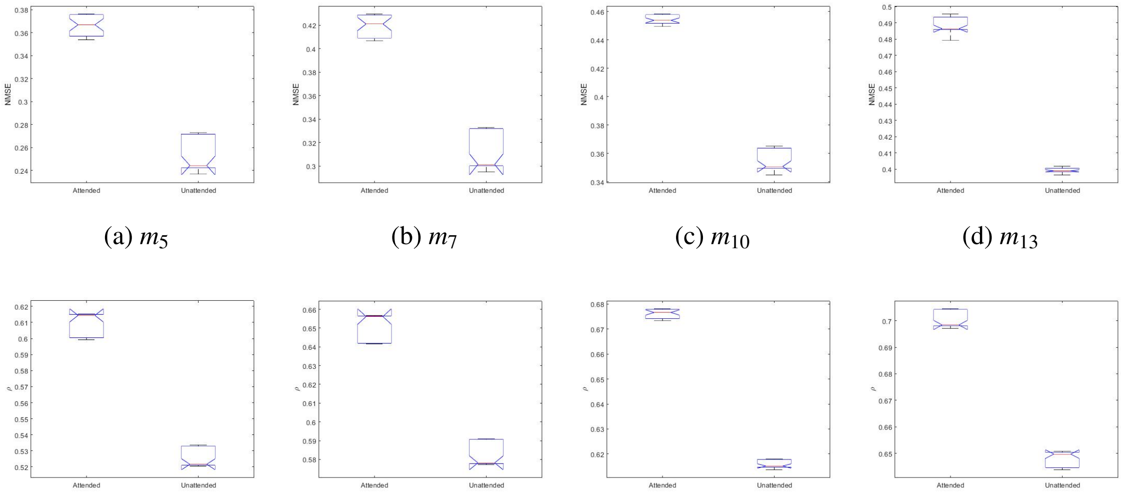
In a similar way, Table 6.2 shows the decoding accuracy (%) for all 30 trials of all subjects based on ρ. Results obtained using this measure were similar to those obtained using NMSE; the maximum classification rates using ρ were also given when using frames with $25ms$ length and for all the selected values of M cepstral coefficients in this initial setup. These decoding accuracies were above 90%. For this $25ms$ frame length, the maximum decoding accuracy (94.73%) was achieved when $M = 5$, i.e. when keeping the first 5 cepstral coefficients (after coefficient m_0) of each frame. The highest mean correlation ($\rho = 0.701$) was obtained when using this same frame length and $M = 13$, i.e. keeping the first 13 cepstral coefficients (after coefficient m_0) of each frame. Table 6.1 shows the mean of ρ (*attended* and *unattended*) and its standard deviation (in parentheses). A Kruskal-Wallis test [12] was performed in order to determine if the obtained values for $\rho_{attended}$ and $\rho_{unattended}$ for all trails of all subjects with frames $25ms$ length were significantly different. The results of this test indicate that ρ values for the *attended* and *unattended* speeches are significantly different (χ^2 = 45.76, p<0.001). Box plots of ρ for all trails of all subjects using frames with $25ms$ length and for each value of M are shown in Figure 6.2. Box plots of ρ for each subject's 30 trials are shown in Figure 6.3.

Table 6.1 Decoding accuracy using NMSE, mean of NMSE and standard deviation (in parentheses) for both *attended* and *unattended* speeches.

Frame length	c_m	Decoding Accuracy (%)	$NMSE_{attended}$		$NMSE_{unattended}$	
25	c_5	94.52	0.362	(0.009)	0.2561	(0.015)
	c_7	92.47	0.4190	(0.01)	0.3152	(0.016)
	c_{10}	95.48	0.4544	(0.003)	0.3560	(0.008)
	c_{13}	96.34	0.4893	(0.005)	0.3993	(0.002)
50	c_5	93.55	0.3513	(0.015)	0.2331	(0.021)
	c_7	91.61	0.3935	(0.016)	0.2838	(0.023)
	c_{10}	94.52	0.4157	(0.012)	0.3052	(0.018)
	c_{13}	53.66	0.3673	(0.049)	0.3318	(0.047)
75	c_5	94.09	0.3515	(0.015)	0.2276	(0.02)
	c_7	91.94	0.3910	(0.016)	0.2771	(0.021)
	c_{10}	54.62	0.3181	(0.054)	0.2824	(0.051)
	c_{13}	53.66	0.3504	(0.047)	0.3262	(0.047)

Table 6.2 Decoding accuracy using ρ, mean of ρ and standard deviation (in parentheses) for both *attended* and *unattended* speeches.

Frame length	c_m	Decoding Accuracy (%)	$\rho_{attended}$		$\rho_{unattended}$	
25	c_5	94.73	0.6078	(0.006)	0.5267	(0.006)
	c_7	91.94	0.6494	(0.007)	0.5841	(0.007)
	c_{10}	93.12	0.6759	(0.002)	0.6161	(0.002)
	c_{13}	92.80	0.7010	(0.003)	0.6475	(0.003)
50	c_5	94.52	0.5946	(0.01)	0.5008	(0.01)
	c_7	91.83	0.6288	(0.011)	0.5516	(0.011)
	c_{10}	93.44	0.6461	(0.008)	0.5713	(0.008)
	c_{13}	53.66	0.6061	(0.039)	0.5896	(0.039)
75	c_5	95.05	0.5946	(0.01)	0.4952	(0.01)
	c_7	91.94	0.6266	(0.01)	0.5441	(0.01)
	c_{10}	53.66	0.5635	(0.044)	0.5426	(0.044)
	c_{13}	53.66	0.5917	(0.04)	0.5834	(0.04)



(a) m_5 (b) m_7 (c) m_{10} (d) m_{13}

(e) m_5 (f) m_7 (g) m_{10} (h) m_{13}

Figure 6.2 Box plots of NMSE (Figures a,b,c, and d) and ρ (Figures e, f, g, and h) for frames with length 25*ms*. All means between *attended* and *unattended* were significantly different (p<0.001).

Of the various parameters that were used for the stimulus-reconstruction model using cepstral coefficients, the setup that performed the best was the one using a frame of 25*ms* length and keeping the first 13 cepstral coefficients (m_1 to m_{13}, excluding the first coefficient m_0). This frame size yielded the highest decoding accuracy based on NMSE and the highest mean of correlations ρ, which agrees with with the common practice within speech analysis of working with frame sizes of 20 to 30*ms* [11]. The choice of this length comes with a trade-off between having more resolution in the frequency domain but with the cost of not meeting assumption for stationarity. The number of cepstral coefficients that were kept (13) that was found to give the best results has been used for experimental purposes and research.

6.2 Using 25ms Frames and 13 Cepstral Coefficients

The next step after deciding the setup for the stimulus-reconstruction using the cepstral coefficients model was to verify if decoding accuracy results and/or NAME or ρ could be improved. This was done by running the model using EEG data from all 128 channels. Results obtained using 128 channels were not significantly different from the ones obtained using 9 channels (see Tables 6.3 and 6.4). The next step was then to run the model with a further reduced number of channels, namely 3 channels. Results and selection of channels are discussed in this section.

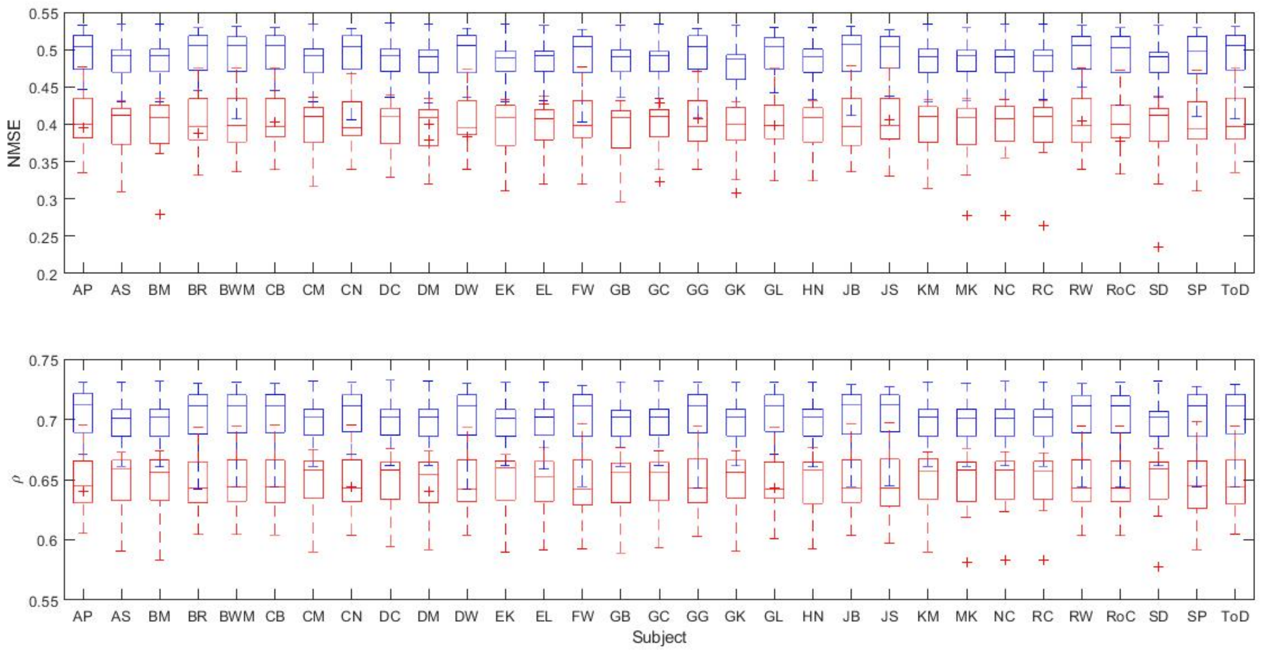
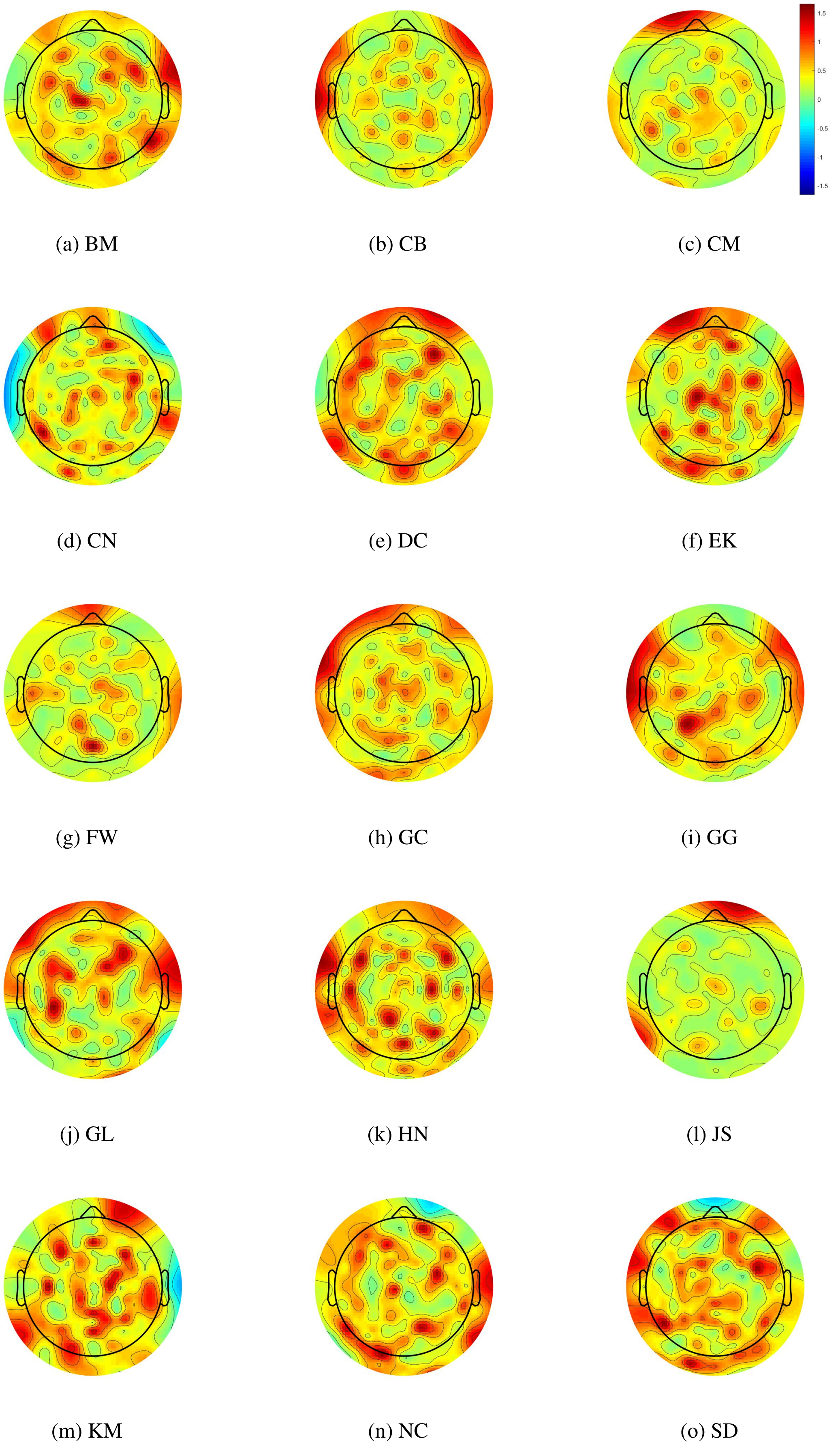


Figure 6.3 Box plot of NMSE and ρ for both *attended* (shown in blue) and *unattended* (shown in red) speeches for the model setup using 9 channels: $A1$, $A19$, $A23$, $B22$, $B26$, $C17$, $C21$, $D18$, $D23$

EEG data from 128 channels were used as *response* for the stimulus-reconstruction using cepstral coefficients model. The decoding accuracy based on NMSE for all trials of all subjects was 96.13%. The mean of NMSE for *attended* was 0.4901, and for *unattended* was 0.4002. The decoding accuracy based on ρ was 93.12%. The mean of ρ for *attended* was 0.7013 and 0.6479 for the *unattended* speech. These results are shown in Table 6.4.

The heat-maps shown in figures Figures 6.4 and 6.5 provide a visualization of the activity of the decoder, **g**, for subjects that attended to speech played to their left and their right ear respectively. From these heat-maps, it is possible to see how the different channels were contributed to the model. For most subjects, the main contribution was given by channels located on top of the head ($A1$) and the ones located at the sides. In this way, a selection of 3 channels was made: $A1$, $B26$, and $D23$. These channels are highlighted in Figure 6.7.



(a) BM

(b) CB

(c) CM

(d) CN

(e) DC

(f) EK

(g) FW

(h) GC

(i) GG

(j) GL

(k) HN

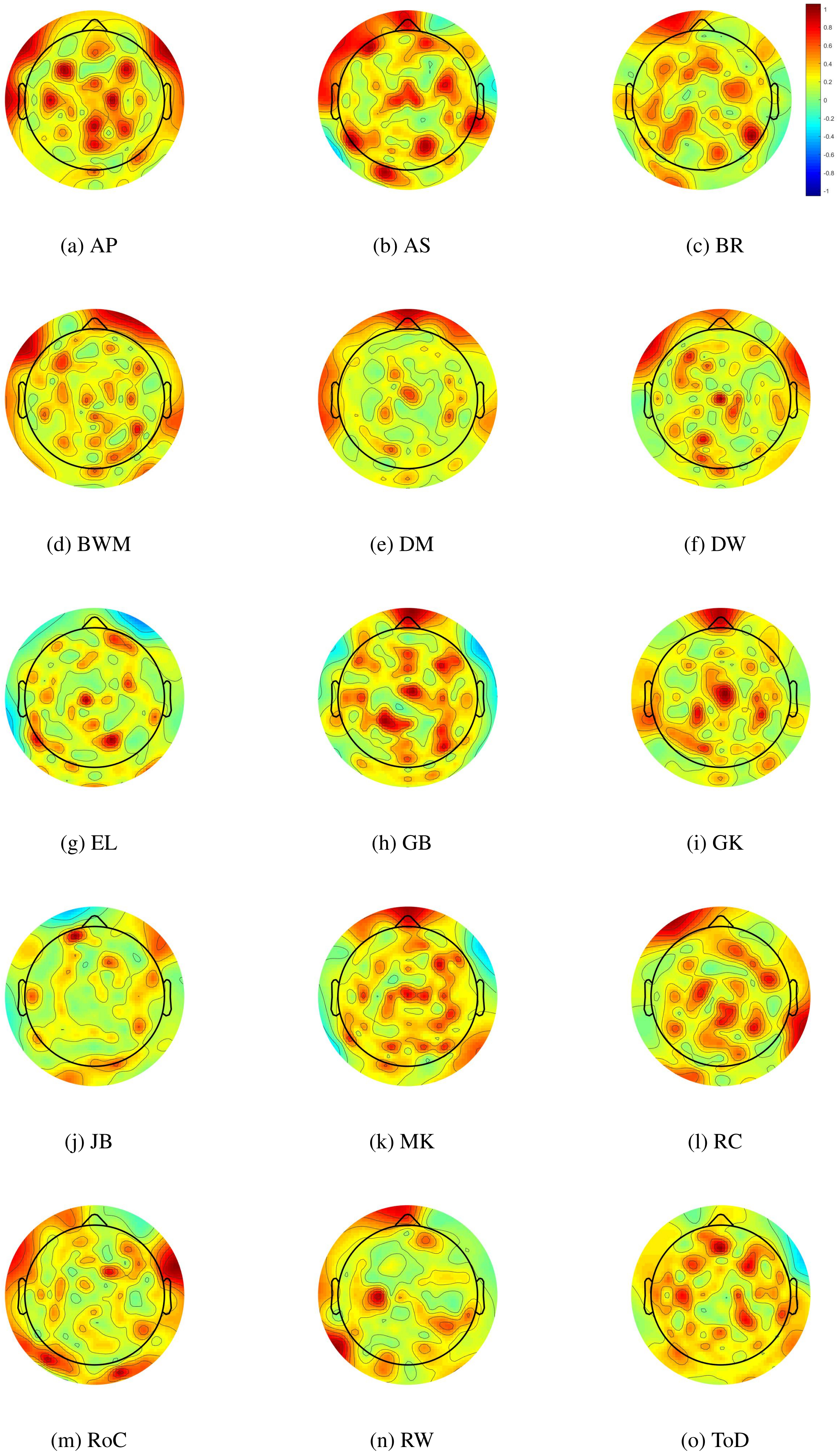
(l) JS

(m) KM

(n) NC

(o) SD

Figure 6.4 Heat-maps of activity of the decoder, normalizzed values of **g**, for subjects who attended to their left ear.



(a) AP

(b) AS

(c) BR

(d) BWM

(e) DM

(f) DW

(g) EL

(h) GB

(i) GK

(j) JB

(k) MK

(l) RC

(m) RoC

(n) RW

(o) ToD

Figure 6.5 Heat-maps of activity of the decoder, normalized values of **g**, for subjects who attended to their right ear.

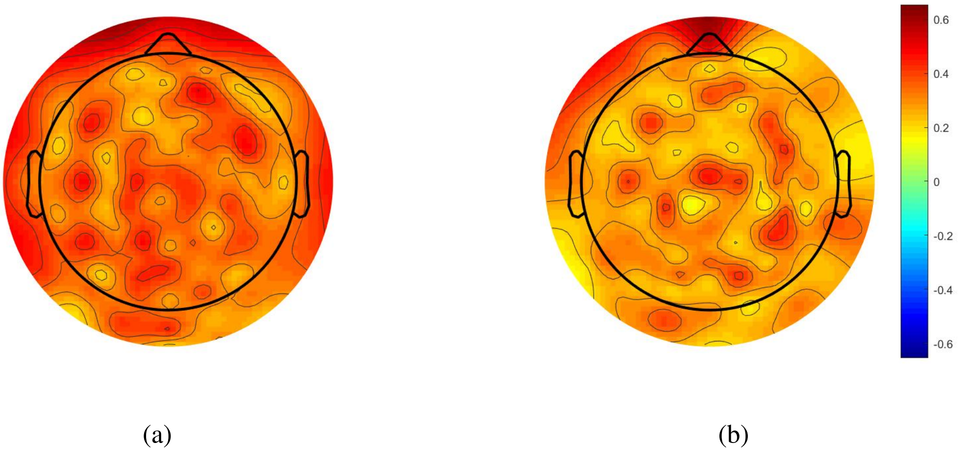


Figure 6.6 Heat-maps of the mean activity decoder, **g** (normalized weights) for subjects who attended to the speech played to their left ear (a) and to the speech attended to their right ear (b).

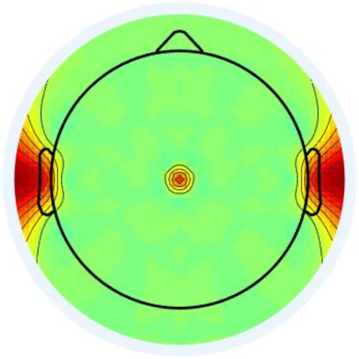


Figure 6.7 Highlighted in red channels $A1$, $B26$, and $D23$.

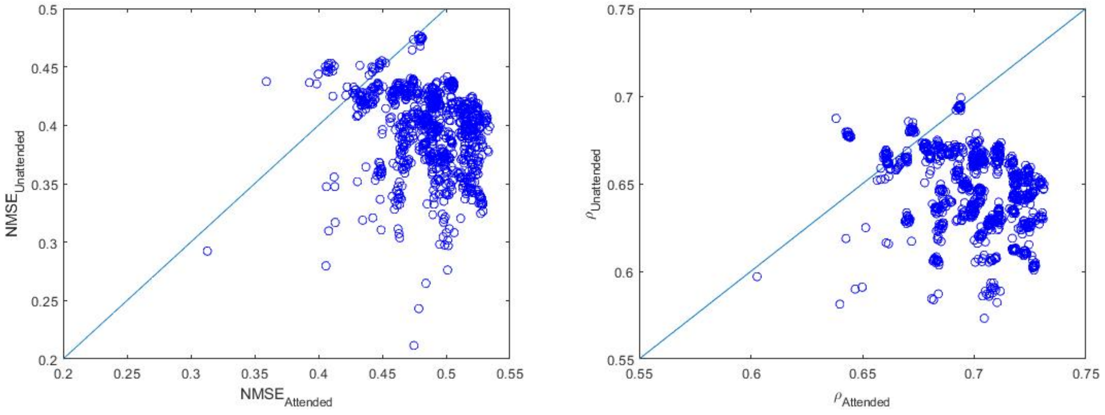
Table 6.3 Decoding accuracy using NMSE, mean of NMSE and standard deviation (in parentheses) for both *attended* and unattended speeches. Result for selected setup using different sets of channels.

Channels	Decoding Accuracy (%)	$NMSE_{attended}$		$NMSE_{unattended}$	
128 channels	96.13	0.4901	(0.005)	0.4002	(0.002)
9 channels	96.34	0.4893	(0.005)	0.3993	(0.002)
3 channels	96.13	0.4870	(0.006)	0.3968	(0.003)

Table 6.4 Decoding accuracy using ρ, mean of ρ and standard deviation (in parentheses) for both attended and unattended speeches. Result for selected setup using different sets of channels.

Channels	Decoding Accuracy (%)	$\rho_{attended}$		$\rho_{unattended}$	
128 channels	93.12	0.7013	(0.003)	0.6479	(0.003)
9 channels	92.80	0.7010	(0.003)	0.6475	(0.003)
3 channels	92.26	0.6996	(0.004)	0.6462	(0.003)

From Figure 6.8 we can see scatter plots for both *NMSE* and ρ obtained from the model using 3 channels for all subjects and all trials. In the $x-axis$ the attended and in the $y-axis$ the unattended. From these plots we can see that the most of the obtained values for both measures, NMSE and ρ using the *attended* speech, are greater than the ones using *unattended* speech. Table 6.5 contains the medians for both *attended* and *unattended* speeches using these 3 channels and the mentioned set up.



(a)

(b)

Figure 6.8 Scatter plots for *NMSE* (a) and ρ (b). In both cases the obtained measures for *attended* were plotted against the *unattended*.

Table 6.5 Median values of NMSE and ρ for both *attended* and *unattended* speeches.

Measure	*attended*	*unattended*
NMSE	0.4855	0.3968
ρ	0.6976	0.6440

6.3 Envelope Vs Cepstral Coefficients as Stimulus

Previous research performed by O'Sullivan et al. [21] was carried out using the envelope of the speech as *stimulus* and the EEG as the *response*. This was made using the stimulus-reconstruction model. The obtained mean of the decoding accuracy for all trials for all subjects was 89.69% achieving correlations between -0.1 and 0.15 [21]. Our results are different from the ones obtained by O'Sullivan et al. [21]. For comparison purposes, Figure 6.9 displays the mean of decoding accuracies of 30 trials for each subject based on NMSE, ρ, and ρ values from that previous study using envelope as *stimulus*. The mean of the decoding accuracy for all trials for all subjects obtained using the stimulus-reconstruction model was 96.13% based on NMSE and 92.25% based on ρ. Though the mean decoding accuracies from both models are high, the ρ values obtained using the envelope as *stimulus* (median $\rho = 0.0054$ [21]) were lower compared to the ones obtained using cepstral coefficients as *stimulus* (median $\rho = 0.6976$).

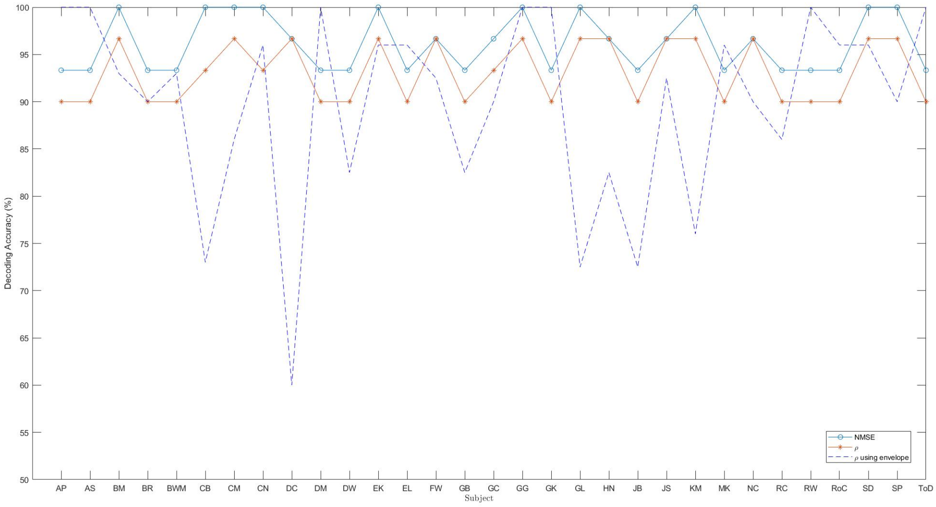


Figure 6.10 Mean of decoding accuracy of 30 trials for each subject based on NMSE (blue line), based on ρ (red line), and based on ρ using the envelope for stimulus-reconstruction (dashed line).

6.4 Preliminary Model Validation

Results obtained showed that under the specifications and conditions of how this experiment was carried out, it is possible to identify to which of the recordings the subjects were trying to attend. The obtained correlations coefficients ρ achieved with the stimulus-reconstruction

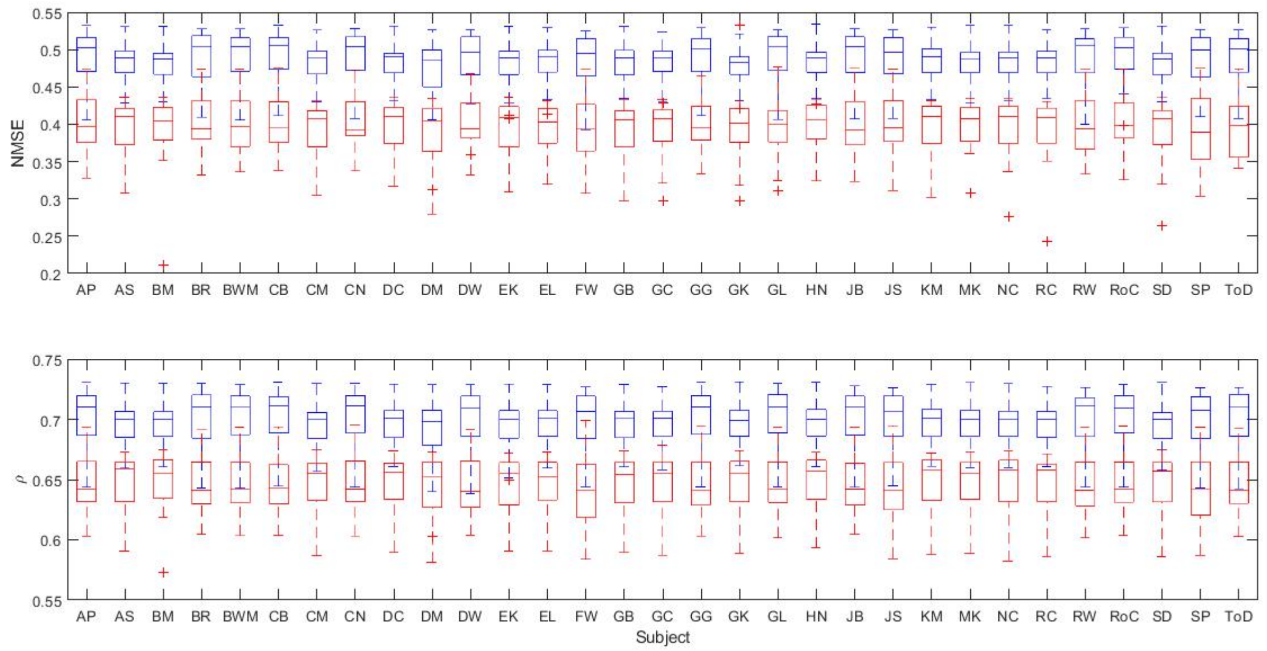
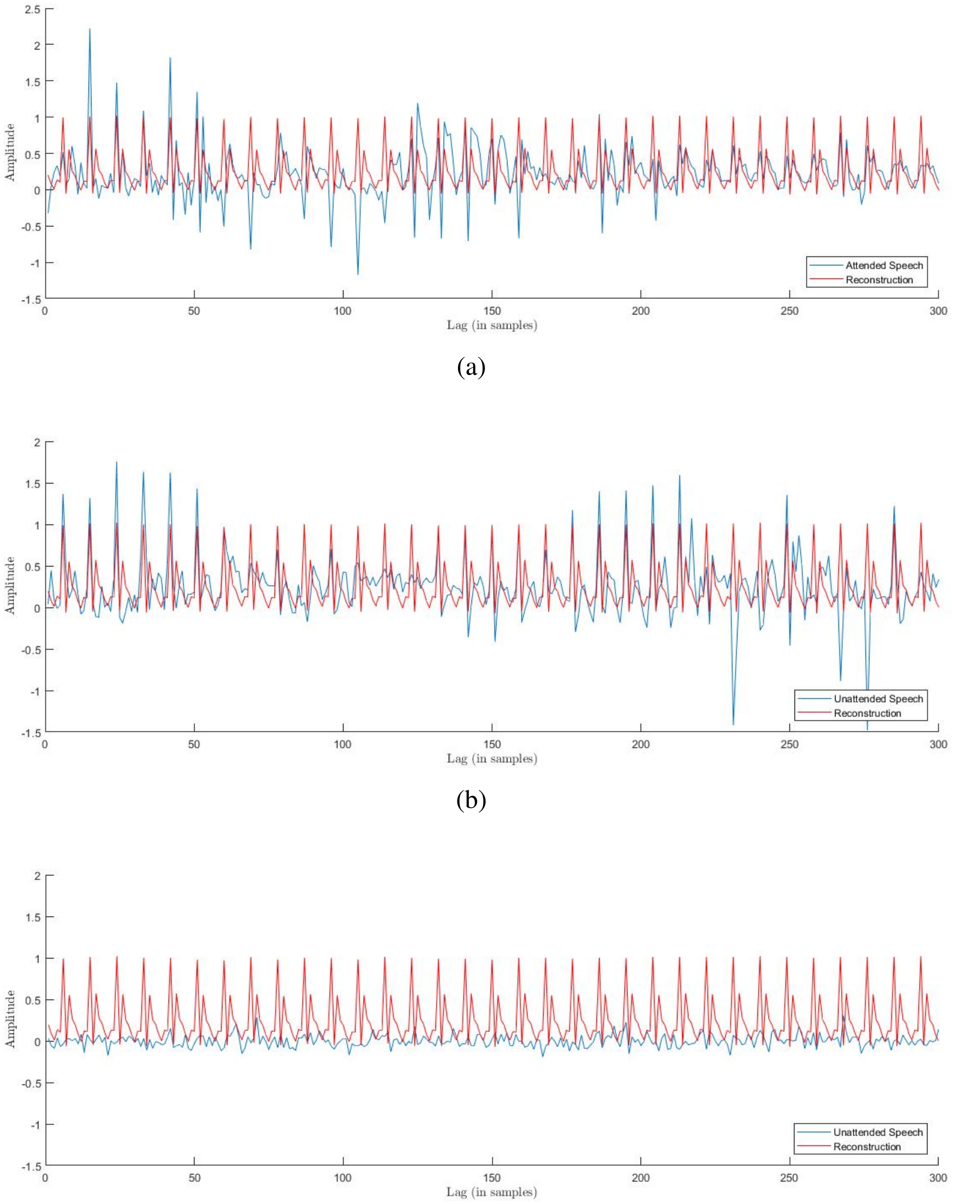


Figure 6.9 Box plot of NMSE and ρ for both *attended* (shown in blue) and *unattended* (shown in red) speeches for the model setup using 3 channels: $A1$, $B26$, and $D23$.

model using cepstral coefficients were considerably higher compared to the ones obtained in previous reports, as the one performed by O'Sullivan et al. [21]. In order to illustrate these reconstructions obtained from the stimulus-reconstruction model using cepstral coefficients; Figure 6.11(a) shows a segment of the cepstral coefficients, $\mathbf{s}_{attended}$ obtained for one of the trials from one of the subjects (shown in blue line). In this same graph, the red line corresponds to the reconstruction $\hat{\mathbf{s}}$. Figure 6.11(b) shows a segment of the cepstral coefficients, $\mathbf{s}_{unattended}$ obtained for this same trial from the same subject (shown in blue line) and the red line corresponds to the reconstruction $\hat{\mathbf{s}}$. In order to test the validity of the stimulus-reconstruction model using cepstral coefficients, an exercise was performed using a sequence of random noise as *unattended* speech. Let $\mathbf{s}_{random}$ be the set of cepstral coefficients from this random sequence. This $\mathbf{s}_{random}$ was compared to the reconstruction $\hat{\mathbf{s}}$ using NMSE and ρ values. The results showed that there was no correlation between $\hat{\mathbf{s}}$ and $\mathbf{s}_{random}$. Figure 6.11(c) shows a segment of the cepstral coefficients $\mathbf{s}_{random}$ obtained from this random sequence, (shown in blue line) and the red line corresponds to the reconstruction $\hat{\mathbf{s}}$.

Figure 6.12 shows the NMSE and ρ values for the first 10 trials from this exercise, Figure 6.12(a) and (b) correspond to the NMSE and ρ respectively that were obtained when using the original *unattended* speech. Figure 6.12(c) and (d) correspond to the NMSE and ρ respectively that were obtained when using a random sequence as *unattended* speech. Similar results were obtained when this same exercise was replicated using data from other subjects.

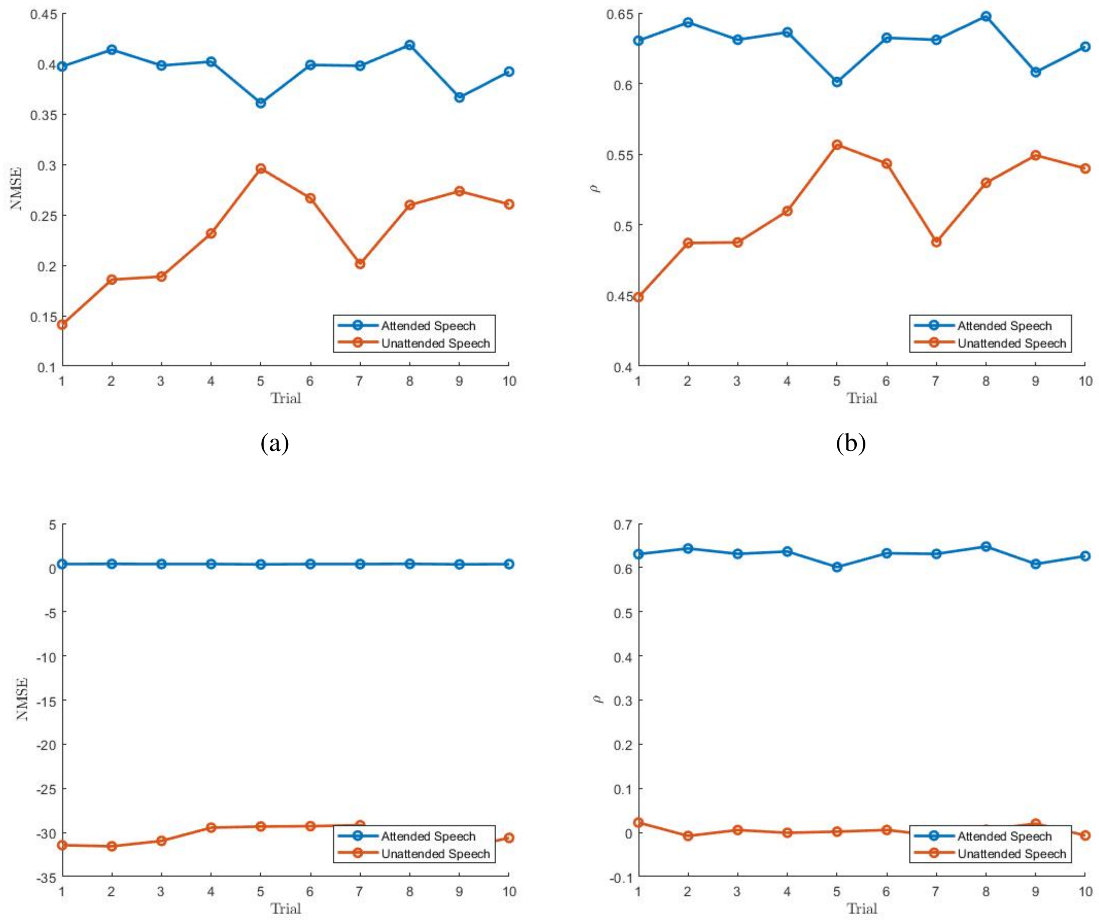


(a)

(b)

(c)

Figure 6.11 Cepstral coefficients $\mathbf{s}_{attended}$ (Figure a), $\mathbf{s}_{unattended}$ (Figure b), and $\mathbf{s}_{random}$ (Figure c) shown in blue. The red line corresponds to the reconstruction $\hat{\mathbf{s}}$.



(a)

(b)

(c)

(d)

Figure 6.12 Figures (a) and (b) show the NSME and ρ values respectively for the first 10 trials from the described exercise when using the original speech streams. Figures (c) and (d) show the NSME and ρ values respectively for the first 10 trials from this same exercise but using a random sequence as the *unattended* speech. In all 4 figures the blue line corresponds to the NMSE and ρ values for the *attended* speech and the red line corresponds to these values for the *unattended* speech.

Following a similar approach we tried using a completely unrelated speech signal as the *unattended* speech. This *external* speech was a recording of a different story, and read by a different speaker, to those the subjects listened to. For this exercise the results were inconsistent as some trials showed correlation between the cepstral coefficients of the *external* speech and the reconstruction $\hat{\mathbf{s}}$, and some did not. This makes it difficult to draw conclusions from this exercise, as the outcomes do not always agree. Considering signals outside the context of the experiment appears to cause problems for the model in some circumstances.

The best way to test the validity of the model would be to apply it on other data sets from similar experimental setups.

6.5 Further Work

The stimulus-reconstruction model using cepstral coefficients described in this project was only tested on one data set. This dataset was obtained under *ideal* conditions (the subjects were wearing headphones and attending to "clean" speeches with no background noise). As mentioned at the end of section 6.4, the stimulus-reconstruction model using cepstral coefficients needs to be tested on more datasets in order to determine the validity of its performance in identifying an attended speech stream. One test could be to apply the model to a dataset where subjects change attention from one speaker to another part-way through the trail. This would give an idea of the ability of the model to adapt to changes in listener attention. A further extension would be to introduce a third speaker into the experiment and ask the subject to periodically switch attention between the three speakers.

The results obtained for this dataset seem to show potential for cepstral analysis to be applied to this task of determining listening attention in a multi-speaker environment. Considering the data in this dataset, when training the models on attended speech streams for each subject and then testing through comparing the predicted speech stream with the attended and unattended speech streams, we were able to obtain high correct classification rates. Also, using white noise as the unattended speech stream to test the validity of the model, we found that classification rates were still high. However, when some unrelated speech signals were taken as the unattended speech stream, the classification rates were not always consistently high. Therefore further investigation is required to find out more about how well this model is performing. Perhaps the most useful next step will be to see how well the model performs when applied to other datasets collected from similar experiments.

Chapter 7

Conclusion

The aim of this project was to investigate whether cepstral processing techniques could be used to identifying listening attention in a multi-speaker environment. Building on the work of O'Sullivan et al. [21], who used the connection between the envelope of speech and the EEG response to determine listening attention, cepstral coefficients were incorporated into a stimulus-reconstruction model to see if the O'Sullivan et al. [21] results could be replicated or even improved using cepstral analysis. Following standard speech processing practices, based on the assumption that speech is stationary over short-time intervals, this model involved breaking corresponding speech and EEG signals into short-time segments over which cepstral coefficients were calculated. These coefficients were used in a stimulus-reconstruction model which, after training a decoder, allowed us to predict the attended speech stream from the EEG signal. This prediction was then compared to the two respective speech streams and the signal which most closely matched the prediction was determined to be the attended speech stream. Using this method we were able to obtain high classification rates of over 90%, suggesting that the method is effective at distinguishing between attended and unattended speech. However, this model was only tested on a single dataset which means that firm conclusions cannot be drawn at this stage. To investigate the validity of the model further, it would need to be tested on other datasets from similar experiments.

Improvements to the model could have been made by including additional features. This model used cepstral coefficients from the low-quefrency region of the cepstrum but it may also be that higher-quefrency regions of the cepstrum could be used, particularly spikes in the cepstrum that indicate voiced speech. Derivatives of the cepstrum are another feature that could improve performance if included in the model.

In summary, having developed a method of determining listening attention by incorporating cepstral processing techniques into a stimulus-reconstruction model, we have shown that, training and testing the model on the O'Sullivan et al. [21] dataset, classification rates of 90% and above can be achieved. Previous research has shown that there is a connection between speech stimuli and the neural response. The results obtained from this project suggest that cepstral analysis, a technique that has proved useful in speech analysis, can be used to find connections between speech and EEG. This potentially adds a new tool for future attempts at determining listening attention in multi-speaker environments. It should be noted however, that although these preliminary findings suggest that cepstral analysis can be used in this problem, further investigation and application of the model to other datasets is required before firm conclusions can be drawn.

Bibliography

[1] Bartlett, M. S. (1948). Smoothing Periodograms from Time-Series with Continuous Spectra. *Nature*, 161(4096):686–687.

[2] Benesty, J., Sondhi, M. M., Huang, Y., and Greenberg, S. (2008). *Springer Handbook of Speech Processing*.

[3] Bogert, B., Healy, M., and Tukey, J. (1963). The quefrency analysis of time series for echos: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking. In Rosenblatt, M., editor, *Proceedings of the Symposium on Time Series Analysis*. Wiley, New York.

[4] Boynton, G. M., Engel, S. A., Glover, G. H., and Heeger, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience*, 16(13):4207–21.

[5] Boynton, G. M., Engel, S. A., and Heeger, D. J. (2012). Linear systems analysis of the fMRI signal. *NeuroImage*, 62(2):975–984.

[6] Bronkhorst, A. W. (2000). BronkhorstCocktail_partyActa_acustica_2000.pdf.

[7] Bronkhorst, A. W. (2015). The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Attention, Perception, and Psychophysics*, 77(5):1465–1487.

[8] Crosse, M. J., Di Liberto, G. M., Bednar, A., and Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *Frontiers in Human Neuroscience*, 10(November):1–14.

[9] Ding, N. and Simon, J. Z. (2012a). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, 109(29):11854–11859.

[10] Ding, N. and Simon, J. Z. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, 107(1):78–89.

[11] Huang, X., Alex, A., and Hsiao-Wuen, H. (2001). Spoken Language Processing.

[12] Kruskal, W. H. and Wallis, W. A. (2016). Use of Ranks in One-Criterion Variance Analysis. 47(260):583–621.

[13] Lindgren, G., Rootzén, H., and Sandsten, M. (2014). *Stationary stochastic processes for scientists and engineers*.

[14] Matlab reference page (2018). *https://www.mathworks.com/help/ident/ref/goodnessoffit.html*.

[15] Mesgarani, N., David, S., Fritz, J., and Shamma, S. (2009). Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. *Journal of neurophysiology*, 102(6):3329.

[16] Mirkovic, B., Debener, S., Jaeger, M., and De Vos, M. (2015). Decoding the attended speech stream with multi-channel EEG: Implications for online, daily-life applications. *Journal of Neural Engineering*, 12(4).

[17] Noll, A. M. (1964). Short-Time Spectrum and "Cepstrum" Techniques for Vocal-Pitch Detection. *The Journal of the Acoustical Society of America*, 36(2):296–302.

[18] Noll, A. M. and Schroeder, M. R. (1964). Short-Time Cepstrum Pitch Detection. *The Journal of the Acoustical Society of America*, 36(5):1030.

[19] Oppenheim, A. V. and Schafer, R. W. (2004). From frequency to quefrency: A history of the cepstrum. *IEEE Signal Processing Magazine*, 21(5):95–100.

[20] Oppenheim, A. V., Schafer, R. W., and Stockham, T. G. (1968). Nonlinear Filtering of Multiplied and Convolved Signals. *Proceedings of the IEEE*, 56(8):1264–1291.

[21] O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., and Lalor, E. C. (2014). Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cerebral Cortex*, 25(7):1697–1706.

[22] 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).

[23] Power, A. J., Foxe, J. J., Forde, E. J., Reilly, R. B., and Lalor, E. C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. *European Journal of Neuroscience*, 35(9):1497–1503.

[24] Proakis, J. G. and Manolakis, D. G. (1996). *Digital Signal Processing*.

[25] Ricketts, T. A. (2005). Directional hearing aids: Then and now. *The Journal of Rehabilitation Research and Development*, 42(4s):113.

[26] Sanei, S. (2013). *Adaptive Processing of Brain Signals*. John Wiley & Sons Ltd.

[27] Schnupp, J., Nelken, I., and King, A. (2011). *Auditory Neuroscience: Making Sense of Sound*, volume 53.

[28] Smith, S. W. (2003). *The Scientist and Engineer's Guide to Digital Signal Processing*.

[29] Stoica, P. and Moses, R. (2009). *Spectral Analysis of Signals*.

[30] Thomson, D. (1982). Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70(9):1055–1096.

[31] Welch, P. (1967). The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms. *IEE Transaction on Audio and Electroacoustics*, AU-15(2):70–73.

[32] Wöstmann, M., Fiedler, L., and Obleser, J. (2017). Tracking the signal, cracking the code: speech and speech comprehension in non-invasive human electrophysiology. *Language, Cognition and Neuroscience*, 32(7):855–869.

Appendix A

Data Set

Participants

Forty human subjects took part (mean ± standard deviation (SD) age, 27.3 ± 3.2 years; 32 male; 7 left-handed). The experiment was undertaken in accordance with the Declaration of Helsinki. The Ethics Committees of the Nathan Kline Institute and the school of Psychology at Trinity College Dublin approved the experimental proceedures and each subject provided written informed consent. Subjects reported no history of hearing impairment or neurological disorder. These data have been published previously using a different analysis approach (Power et al. [23], O'Sullivan et al. [21]).

Stimuli and Procedures

Subjects undertook 30 trials, each of approximately 1 minute in length, where they were presented with 2 classic works of fiction: one to the left ear, and the other to the right ear. Each story was read by a different male speaker. Subjects were divided into two groups of 20 with each group instructed to attend to the story in either the left or right ear throughout all 30 trials. After each trial, subjects were required to answer between 4 and 6 multiple-choice questions on both stories. Each question had 4 possible answers. We used a between-subjects design as we wanted each subject to follow just one story to make the experiment as natural as possible and because we wished to avoid any repeated presentation of stimuli. For both stories, each trial began where the story ended on the previous trial. Stimulus amplitudes in each audio stream within each trial were normalised to have the same root mean squared (RMS) intensity. In order to minimise the possibility of the unattended stream capturing the subjects' attention during the silent periods in the attended stream, silent gaps exceeding 0.5 s were truncated to 0.5 s in duration. Stimuli were presented using Sennheiser HD650 headphones and Presentation software from Neurobehavioral Systems (http://www.neurobs.com). Subjects were instructed to maintain visual fixation for the duration of each trial on a crosshair centred on the screen, and to minimise eye blinking and all other motor activities.

Data Acquisition and Preprocessing

Electroencephalography data were recorded for 34 of the subjects (17 of these subjects attended to the speech on the left and the remaining 17 to the right) using 128 electrode positions (shown in fig. A.1). Data for the remaining 6 participants were collected using 160 electrode positions (3 of these subjects attended to the left and the remaining 3 to the right). These data were then remapped to an equivalent 128 electrode positions using an interpolated spline function. The data were filtered over the range 0-134 Hz and digitised at the rate of 512Hz using a BioSemi Active Two system. Data were referenced tot he average of all scalp channels.

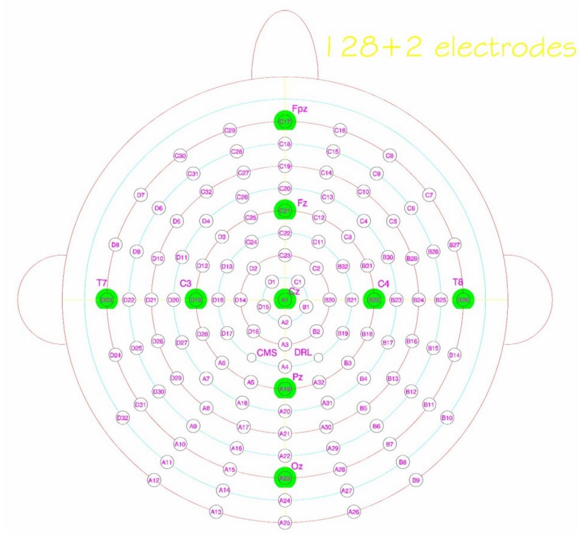


Figure A.1 EEG electrode layout known as "biosemi layout 128". This was the layout used to collect the data considered in this study.