Institutionen för systemteknik
Department of Electrical Engineering

Examensarbete

Sparse Linear Modeling of Speech from EEG

Examensarbete utfört i Reglerteknik
vid Tekniska högskolan vid Linköpings universitet
av

Mattias Tiger

LiTH-ISY-EX-ET–14/0420–SE

Linköping 2014

Department of Electrical Engineering
Linköpings universitet
SE-581 83 Linköping, Sweden

Linköpings tekniska högskola
Linköpings universitet
581 83 Linköping

Sparse Linear Modeling of Speech from EEG

Examensarbete utfört i Reglerteknik
vid Tekniska högskolan vid Linköpings universitet
av

Mattias Tiger

LiTH-ISY-EX-ET–14/0420–SE

Handledare: **Johan Dahlin**
ISY, Linköping University
Eline Borch Petersen
Eriksholm Research Center, Oticon
Thomas Lunner
Eriksholm Research Center, Oticon

Examinator: **Thomas Schön**
ISY, Linköping University

Linköping, 9 juni 2014

Till min älskade pappa

Avdelning, Institution
Division, Department

Automatic Control
Department of Electrical Engineering
SE-581 83 Linköping

Datum
Date

2014-006-09

Språk
Language

□ Svenska/Swedish
☒ Engelska/English
□ _____

Rapporttyp
Report category

□ Licentiatavhandling
☒ Examensarbete
□ C-uppsats
□ D-uppsats
□ Övrig rapport
□ _____

ISBN
—

ISRN
LiTH-ISY-EX-ET–14/0420–SE

Serietitel och serienummer
Title of series, numbering

ISSN
—

URL för elektronisk version
http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-108048

Titel
Title

Gles Linjära Modellering av Tal från EEG
Sparse Linear Modeling of Speech from EEG

Författare
Author

Mattias Tiger

Sammanfattning
Abstract

For people with hearing impairments, attending to a single speaker in a multi-talker background can be very difficult and something which the current hearing aids can barely help with. Recent studies have shown that the audio stream a human focuses on can be found among the surrounding audio streams, using EEG and linear models. With this rises the possibility of using EEG to unconsciously control future hearing aids such that the attuned sounds get enhanced, while the rest are damped. For such hearing aids to be useful for every day usage it better be using something other than a motion sensitive, precisely placed EEG cap. This could possibly be archived by placing the electrodes together with the hearing aid in the ear.

One of the leading hearing aid manufacturer Oticon and its research lab Erikholm Research Center have recorded an EEG data set of people listening to sentences and in which electrodes were placed in and closely around the ears. We have analyzed the data set by applying a range of signal processing approaches, mainly in the context of audio estimation from EEG. Two different types of linear sparse models based on L1-regularized least squares are formulated and evaluated, providing automatic dimensionality reduction in that they significantly reduce the number of channels needed. The first model is based on linear combinations of spectrograms and the second is based on linear temporal filtering. We have investigated the usefulness of the in-ear electrodes and found some positive indications. All models explored consider the in-ear electrodes to be the most important, or among the more important, of the 128 electrodes in the EEG cap. This could be a positive indication of the future possibility of using only electrodes in the ears for future hearing aids.

Nyckelord
Keywords

EEG, In-Ear, Sparse, L1-regularization, Least Squares, FIR, Spectrogram, Machine Learning

Abstract

For people with hearing impairments, attending to a single speaker in a multi-talker background can be very difficult and something which the current hearing aids can barely help with. Recent studies have shown that the audio stream a human focuses on can be found among the surrounding audio streams, using EEG and linear models. With this rises the possibility of using EEG to unconsciously control future hearing aids such that the attuned sounds get enhanced, while the rest are damped. For such hearing aids to be useful for every day usage it better be using something other than a motion sensitive, precisely placed EEG cap. This could possibly be archived by placing the electrodes together with the hearing aid in the ear.

One of the leading hearing aid manufacturer Oticon and its research lab Erikholm Research Center have recorded an EEG data set of people listening to sentences and in which electrodes were placed in and closely around the ears. We have analyzed the data set by applying a range of signal processing approaches, mainly in the context of audio estimation from EEG. Two different types of linear sparse models based on L1-regularized least squares are formulated and evaluated, providing automatic dimensionality reduction in that they significantly reduce the number of channels needed. The first model is based on linear combinations of spectrograms and the second is based on linear temporal filtering. We have investigated the usefulness of the in-ear electrodes and found some positive indications. All models explored consider the in-ear electrodes to be the most important, or among the more important, of the 128 electrodes in the EEG cap. This could be a positive indication of the future possibility of using only electrodes in the ears for future hearing aids.

Acknowledgments

As a final part of the Ph.D course in Machine Learning by Prof. Thomas Schön, we were encouraged to do a project. I had multiple extra courses at the time and we agreed that I would be able to postpone it till the summer. I wasn't sure what to do, but Prof. Schön had an interesting idea. They work together with the research department of Danish Oticon to improve the signal processing in hearing aids. An at the time recent publication in Nature by another research group had published very positive results regarding using EEG to distinguish what sound a subject is focused on. A prospect is that this might be possible to use in future hearing aids. The idea was to do some research work in this direction. It got me interested and excited as my father is suffering from hearing impairment and uses hearing aids. Although it helps him as a professional musician, he sometimes has trouble following conversations with multiple speakers and usually gets very tired from all the additional concentration needed. After some discussions the proposed project turned into a summer-job and this thesis. I hope that I have been able to contribute, even if ever so slightly, towards the goal of creating better hearing aids in the future.

I want to thank my examiner Prof. Thomas Schön for giving me this opportunity, and for giving such great courses and explaining difficult concepts with remarkable clarity. Also I want to thank Oticon, Eriksholm Research Center and specially Dr. Thomas Lunner extensively for this opportunity, and for very interesting discussions and insights into a fascinating domain. Furthermore I want to thank my supervisor Johan Dahlin and co-supervisor Eline Borch Petersen whom provided great guidance, feedback and support when writing this thesis. I greatly appreciate your support and enlightening discussions concerning knowledge from your respective fields, which to a large extent made this thesis possible. A special thanks to Fredrik Gustavsson and Sina Khoshfetrat Pakazad for their valuable advice and discussions as well as Svante Gunnarsson and everyone at the Automatic Control department at ISY for a really great time. You gave me an opportunity to work on my thesis while simultaneously giving me a preview of the life of a Ph.D student in a warm and friendly atmosphere.

Finally I want to thank my family and especially my beloved Hanna Johansson for all their support during my years of studying and the writing of this thesis.

Linköping, Juni 2014
Mattias Tiger

Contents

1 Introduction

1.1 Background

Humans possess a remarkable ability to attend to a single speaker's voice in a multi-talker background. How the auditory system manages to extract intelligible speech under such acoustically complex and adverse listening conditions is not known and it is not clear how the attended speech is internally represented (Mesgarani & Chang, 2012). What is known is that many social situations throughout life require this ability to attend single sound sources in the surroundings of many, and that a lesser such ability hamper ones engagement and interactions with other people. People with impaired hearing capabilities often have a great difficulty with attending a single speaker's voice in a multi-talker background, even with the usage of modern hearing aid. This problem threatens the participation of social gatherings or even discussions over the dinner table and an improvement of this ability might provide hearing impaired people with a noticeable higher quality of life. The phenomenon of being able to focus one's auditory attention on a particular stimulus while filtering out other stimuli is known as the cocktail party effect (A. Bronkhorst, 2000), and the problem of how to automatically do this is commonly known as the cocktail party problem.

The largest hearing aid manufacturer Oticon and their research center Erikholm Research Center are interested in investigating the possibility of improving this condition with future hearing aids. Recent studies have shown that the audio stream a human focuses on can be detected among the surrounding audio streams, using EEG and linear models. With this rises the possibility of using EEG to unconsciously control future hearing aids such that the attuned sounds get enhanced, while the rest are damped. For such hearing aids to be useful for every day usage it better be using something other than a motion sensitive, precisely

placed EEG cap. This could possibly be archived by placing the electrodes together with the hearing aid in the ear.

1.2 Related work

It has been shown that the audio stream focused on can be found among all heard audio streams using EEG, this using linear models. Mesgarani & Chang (2012) have in a recent study shown the possibility of using EEG to detect which sound source a human attends to. This is possible since a part of the EEG from early in the brain's audio processing have a strong correlation to the spectrogram of the attended sound source, and low correlation to the rest of the surrounding sound. In the study the EEG is measured using sensors placed directly on the brain tissue, using a high resolution grid of sensors. Lalor et. al. (2012) builds upon this work and explores three different decoding approaches: one based on Canonical Correlation Analysis (CCA), one based on single channel inversion (AESPA) and one based on all channel inversion, where the last one produces attention decoding accuracy of 75-95% with 60 seconds of data. All three approaches are over 90% accurate in speech decoding in a 20 ms frame. The all channel inversion is performed by an estimated multivariate linear filter. Rajaram, Lalor & Shinn-Cunningham (2013) uses CCA to find an optimal EEG channel subspace and lag, then decoders are trained for attended streams, unattended streams and mixtures of such, with respect to maximizing the respective correlation. Classification of the attended sound source is 80% accurate.

1.3 Contribution

Erikholm Research Center have recorded a EEG data set from people listening to sentences and in which some electrodes were placed in and closely around the ears. We have analyzed this data set by applying a range of signal processing approaches, mainly in the context of audio envelope estimation from EEG. Two different types of linear sparse models based on L1-regularized least squares are formulated and evaluated, providing automatic dimensionality reduction by significantly reducing the number of channels used. The first model is based on linear combinations of spectrograms and the second is based on linear temporal filtering, both are compared to CCA. Lag estimation using CCA is evaluated on the data set. The data sets contain measurements from electrodes placed in the ears of the participants, and the usefulness of these electrodes are explored.

1.4 Result

The two sparse models are sufficiently flexible to explain the transformation from EEG to sound envelope using few channels. It is not clear from this study that the models can generalize to new data not trained on. They are unable to do this given the data set in the study, possibly due to too short data sequences or lack-

ing sufficient model dynamics. Improvements and error sources are discussed. The sparse models do however provide insight into the number of channels necessary to reconstruct the audio envelope from EEG data. Dimension reduction of the EEG signal is performed by automatic electrode selection using these two different sparse models. Most channels can be removed this way without significantly decreeing the estimate's quality. Using the very dynamic temporal filtering model fewer than seven electrodes are necessary to explain single sentences to a very high degree from EEG data. Positive indications of the usefulness of having electrodes in the ears for measuring auditory EEG responses is seen. Electrodes put in ears are among the most important in the two different sparse model's estimates. This could be a positive indication of the future possibility of using only electrodes in the ears for future hearing aids. Efficient sparse model training is possible: The temporal filter model is reformulated to the frequency domain, which together with a well structured sparse matrix improve the time it takes to solve it significantly. Many avenues for further work exists, some of which are presented throughout the report.

1.5 Outline

The first two chapters introduce the data set used in the study and present an analysis using traditional signal processing techniques. Chapter four and five cover the formulation and evaluation of the two explored sparse models. A comparison between the models and a discussion is presented in chapter six.

Chapter 2 provides background about EEG signals and the data set. Properties of the EEG, noise sources and the performed pre-processing and noise reduction is described.

Chapter 3 contain a preliminary data analysis consisting of some standard signal processing methods. The methods intend to maximize the correlation in increasingly sophisticated ways. Initially using temporal cross-correlation. Then using coherence (correlation in the frequency domain) and finally using CCA which is a affine transform-invariant method that finds the linear basis for respective signal which maximizes the cross-correlation.

Chapter 4 explores a sparse linear model based on spectrograms. An implementation is trained on the data and evaluated.

Chapter 5 explores a sparse linear model based on temporal filtering. An implementation in the form of a FIR filter is trained on the data and evaluated.

Chapter 6 compares the different methods, discusses the results and relate to future work.

Appendix A presents the primarily statistics used to evaluate the models.

2

EEG signals

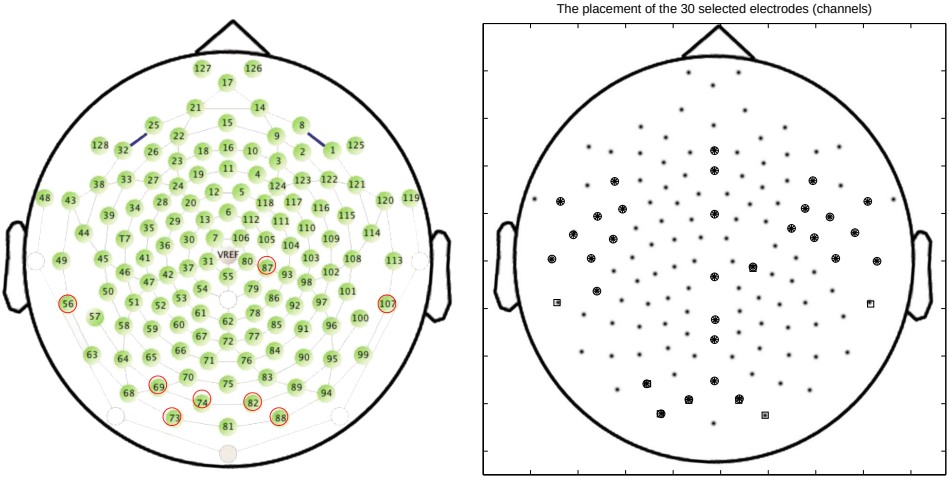
Electroencephalography (EEG) is the recording of electrical fields on the head, with the purpose of capturing brain activity giving rise to the fields. The measurements can be obtained by placing an electrode cap on a subjects scalp or by placing a denser, finer electrode grid directly on the brain. In the former case, the EEG is low pass filtered by the cerebrospinal fluid, the tissue and bone between the brain and the electrode, resulting in the loss of high frequency information in the EEG recording. Apart from the fields resulting from the brain activity under study, natural leakage from other brain processes will be recorded as well. Sensors on the scalp are also susceptible to additional noise sources in the form of muscle contractions, especially eye blinking. The electric fields emitting from the brain are also affected by the tissue in general as well as conductivity between the sensor and the skin. Conductive gel is therefore used to mitigate this problem.

It is known that the frequencies between 2 Hz and 16 Hz in EEG contribute most to speech intelligibility (Powel et. al. 2012). This is due to the direct connection with the mouth gesticulation performed when forming sounds and words which is also typically in this frequency span.

Traditionally most models using EEG are designed for Event-Related Potentials (ERP). ERP is a stereotyped EEG response to a specific stimuli, e.g. a sensory event. ERP is calculated as a time-locked average of EEG from many trials involving the identical stimuli. This is done to improve the signal to noise ratio of the otherwise very noise EEG measurements. Significant events seen in such grand averaging due to auditory stimuli is N100 and often followed by P200, where N100 is a minima at around 100 ms and P200 is a maxima at around 200 ms after the stimuli. Although the data used is insufficient to calculate an ERP, N100 and P200 will be used as a guide to expected responses seen in the EEG.

The data used in this project was recorded by Erikholm Research Center and consists of a passive listening trial. A 128 channel EEG net from EGI was placed on heads as shown in figure 2.1(a). It was used to record the data with a sampling frequency of 250 Hz. One sensor was placed on each ear lobe, three sensors were placed in each ear channel and the rest on their standard position on the head.

The trials consist of passive listening sessions. The test subjects were three normal hearing students (1 male, 2 females). The test subjects listens to a single sentence at a time, without responding. First 50 sentences spoken by a male, then 50 sentences spoken by a female are presented. Pauses between sentences are randomized to be 0.2 ± 0.05 s. Randomizing the pause length is done to compensate for that the brain adapts very well to repetitions. The sentences varied in length with each being around 2.8 seconds on average.



(a) *Electrode positions* ***(b)*** *30 selected channels*

Figure 2.1: ***(a)*** *Electrode position of the EGI EEG net. Electrode 17 is placed between the eyes and 81 in the back of the neck. The the electrodes marked with a red circle are placed in the ear or on its lobe. The sensors are placed more around the head than what the illustrated figure shows.* ***(b)*** *The 30 selected electrodes are marked with stars. The points in a square are the electrodes in the ear, except the two close to the figures ears, on either side, which are placed on respective ear lobe. The rectangle at the bottom of the image with no star is electrode 88.*

A set of 30 channels were selected as probably the most suitable sensors for this task, seen in figure 2.1(b) and table 2.1. Among these are some that are located straight over the head, to capture signals from most of cortex and frontal cortex in particular. Other sensors included are located around the ears, covering the areas over and around audio cortex as well as a few that were inserted into the ears. The sensors at the front of the head usually contain a lot of noise from eye blinking and muscle movement and was therefore not selected.

Channels	**Description**
6, 11, 16, 55, 62, 72, 75	Straight over the head from front to back
69, 73, 74, 82, 87	In the ear
33, 34, 39, 40, 43, 44, 45, 49, 50, 108, 109, 110, 113, 114, 115, 116, 120, 122	Around the ear

Table 2.1: *The 30 selected channels.*

Sensor 88 was initially not considered due to its high impedance, and was therefore not chosen as one of the 30 selected channels. Impedance is in this context a measure of the inversion of how good the connectivity is between the sensor and the head. It is generally preferred that the impedance is below 50 kOhm, but it is not a necessity in order for the signal to be useful. Most of the channels among the 30 selected have an impedance below 50 kOhm. Sensor 88 from the ear had an impedance of over 1500 kOhm for all three persons.

Figure 2.2 illustrates a typical sentence and some of its EEG channels. In this figure, channel 7 show a linear trend contrary to the other channels listed. The trend seen is called baseline drift and is most likely due to increased sweat or due to decreased connectivity between the skin and the electrode. In channels 59 and 72, a build-up over time of increasing variance is seen. Channels 17, 82 and 114 are quite similar and the few dissimilarities could be due to noise. Channels 82 and 88 are very similar as well while in comparison to these two, channels 45 and 51 have a few peaks that are slightly larger and smaller, respectively. Minor high frequency noise can be clearly seen in the channels below, and all channels contain such noise.

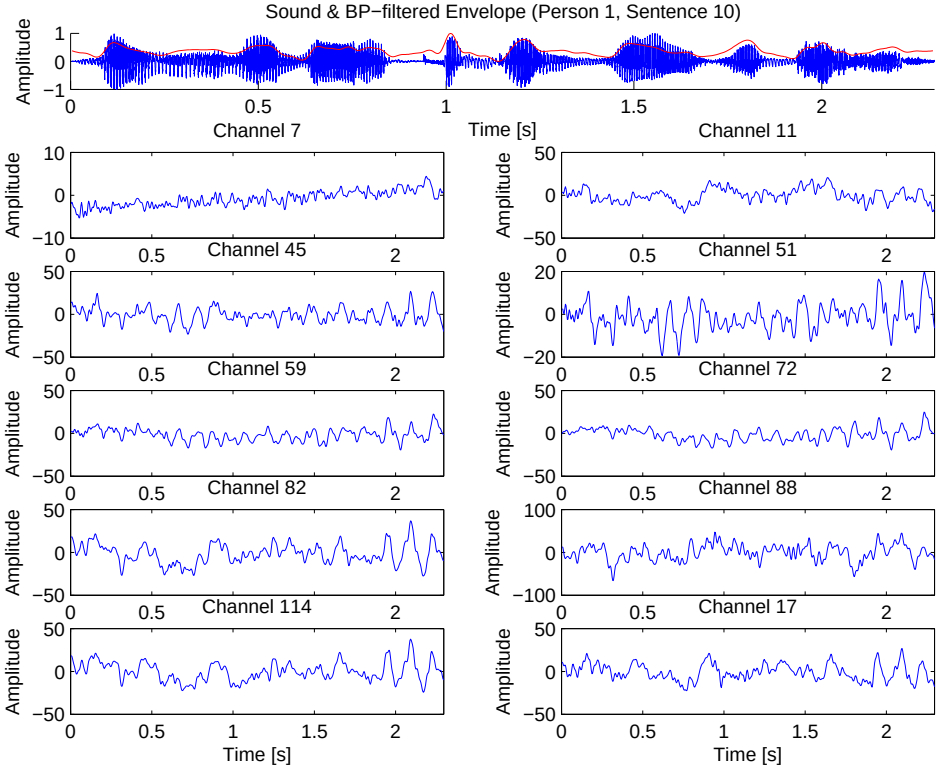


Figure 2.2: *The sound of sentence 10 of person 1 and corresponding EEG signals from a selection of channels. In the top image a BP-filtered sound envelope is indicated by red. The spatial placement of the channels can be seen in figure 2.1(b).*

Some noise is removed with frequency selective filtering according to the following procedure. The sound envelopes are calculated and then down sampled using Real Mean Square (RMS) averaging. The down sampled envelope and the EEG signal are then Band Pass (BP) filtered by a Butterworth filter of order 5 to be within the frequency range of intelligible speech, which is between about 2-16 Hz. In figure 2.3 this is demonstrated using both a BP filter and a Low Pass (LP) filter for comparison. Also shown in the figure is the $\log_{10}$ of the filtered envelope with the renormalization to be between 0 and its maximum from before applying the log function, following roughly the same pre-processing as Power et. al. (2012). The $\log_{10}$ is used in the L1-linear filter model covered in chapter 5 where as in the rest of the report filtered envelopes without taking the $\log_{10}$ are used since no major difference was seen.

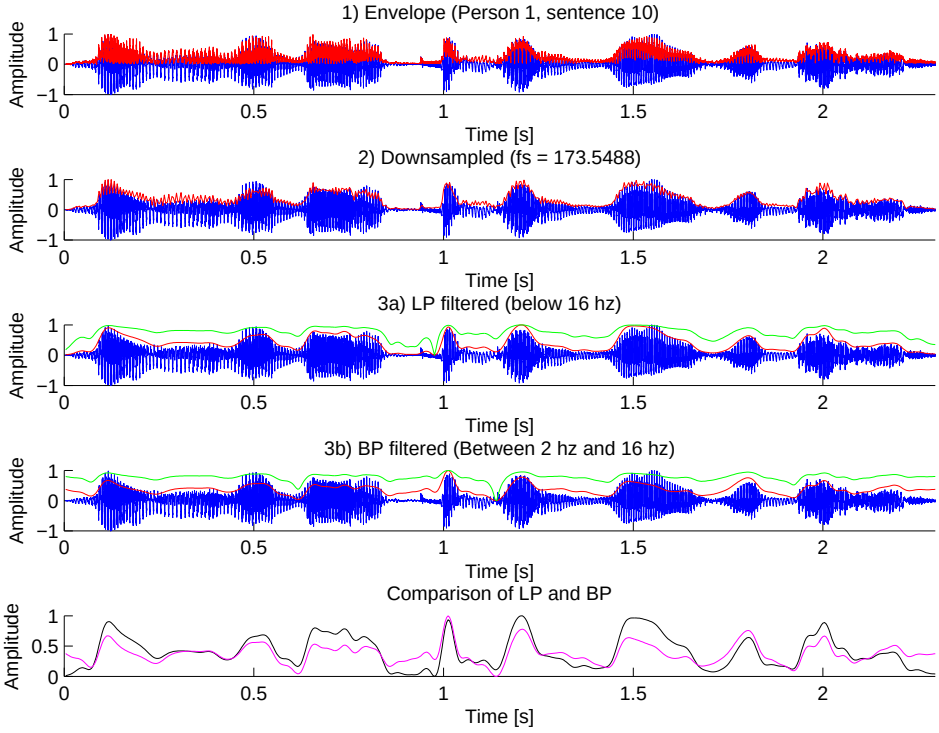


Figure 2.3: *The envelope at different stages of the preprocessing. In the first 4 figures the red signal is the envelope and the green signal is the log10 envelope. The BP filtered envelope has had the mean value added to it, to illustrate the difference to the LP version more clearly. In the figure at the bottom the LP filtered signal is black and the BP filtered is magenta colored.*

3

Preliminary data analysis

A preliminary data analysis is performed to get a better picture of the data. We try out some standard signal processing approaches to see if they show any windows of opportunities for more advanced methods. A search for channels that may correlate more with the sound envelope than other channels is performed. We are interested to see if there are any frequency bands that correlate more to the targeted data. Any correlations between frequency components of the channels and the sound envelope are studied briefly. Finally, a model using Canonical Correlation Analysis is estimated and evaluated. Ultimately a model that at least crudely describe the transformation from EEG to sound envelope is desired, but highly unrealistic due to the properties of the data such as the amount of noise and possibly non-stationarity.

3.1 Correlation analysis

Cross-correlation is a measure of the similarity of two signals as a function of the time-lag. Consequently one interesting aspect to evaluate is if any EEG channel is similar enough to the sound envelope so that it alone can be used to estimate the sound envelope. Estimating the time lag between the sound playing and its first effects on the EEG is also of interest in itself.

Sampled correlation requires a stationary process in order to converge to a correct estimate. The EEG signal could be non-stationary due to transients in the form of gradual change in impedance of the sensors as the electrodes dry, or due to the effects of other mental processes in the background. During the remainder of this section the stationarity of the EEG is assumed, as it is analyzed by this classic signal processing method. In the case of non-stationarity, a more elaborate approach might be offered in Parra & Spence (2000) or in Podobnik & Stanley (2007).

The correlation between the sound envelope and the individual EEG channels are calculated using different lags, denoted τ. These lags simulate the delay between the audio playback and the EEG due to the time it takes the audio to propagate through the ear and the brain processes. Correlation is measured as an absolute value, so in a sense the possible flip of sign is the transformation model used in this section. The cross-correlation is estimated by the expression

$$R_{x_c y}[\tau] = \frac{\sum_k x_c[k] y[k-\tau]}{\sqrt{\sum_k x_c^2[k] \sum_k y^2[k]}}, \tag{3.1}$$

where $x_c[k]$ denotes the value of channel c of an EEG signal at time k and y denotes the corresponding sound envelope. Here, τ is an integer between 0 and N, where N is the length of y. τ is a positive integer since the brain, as any natural system, is causal. Since the data is limited, the cross-correlation estimate for lags other than 0 is calculated using zero-padding, denoted corr in the figures. The EEG signal is then zero-padded at the beginning, simulating the time lag, and the audio signal is zero-padded at the end so that the two signals are of the same length. An alternative way is to truncate the shifted signals so that no zero-padding is used in the correlation estimate for any single lag, this is denoted xcorr. While corr may underestimate the correlation slightly, xcorr is likely to overestimate the correlation since it is not weighted in accordance to the actual number of samples. This is shown in figure 3.3.

For each sentence, the EEG channel and lag recorded that correlate the most to the audio envelope is found. This is accomplished by calculating the cross-correlation between each separate EEG channel and the audio envelope. The lag which provides the maximum correlation to the audio envelope is determined for each channel . The channel which correlates the most is then selected as the best channel and its lag is considered the estimated lag.

The results are exemplified in figure 3.1 and in figure 3.2, illustrating sentence 8 and 35 respectively. Sentence 8 has slightly higher maximum correlation than the mean of all sentences, as seen in figure 3.3, where as sentence 35 is one of the channels with highest maximum correlation. This means that sentence 35 is one of those sentences which have an EEG channel that is more similar (in the sense of correlation) to its audio envelope, compared to any other such pair for most other sentence.

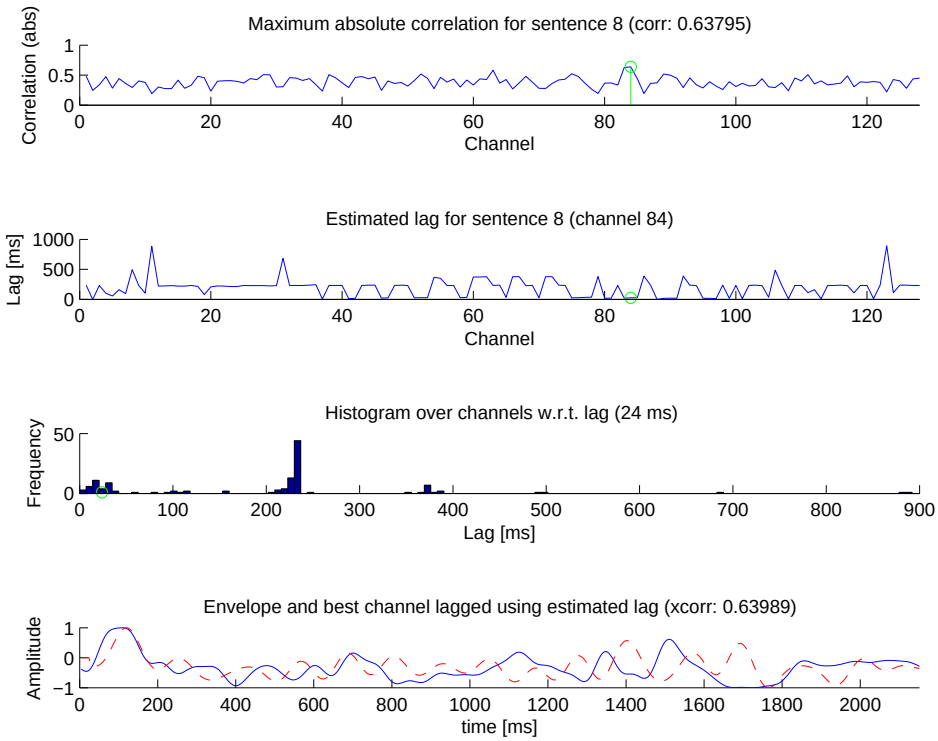


Figure 3.1: *The channel with highest correlation and the estimated lag is marked with a green circle. In the image at the bottom the blue signal is the sound envelope and the red signal is the green-marked channel with the estimated lag and zero-padded to be as long as the sound envelope. (person 1, sentence 8)*

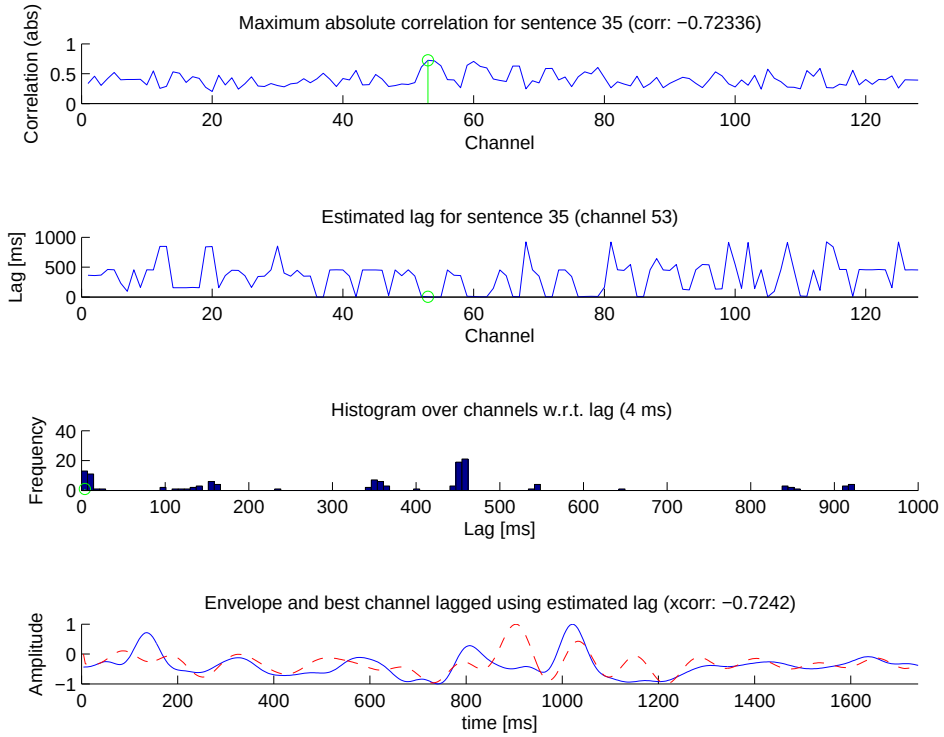


Figure 3.2: *The channel with highest correlation and the estimated lag is marked with a green circle. In the image at the bottom the blue signal is the sound envelope and the red signal is the green-marked channel with the estimated lag and zero-padded to be as long as the sound envelope. (person 1, sentence 35)*

The mean maximal correlation over all sentences is 0.55. Histograms in figure 3.3 show that the estimated lags are concentrated below 200 ms and that the most of them are below 400 ms, which is reasonable. It can also be seen that there presumably are some channels that are less likely to be important than others. As seen in the figure, corr and xcorr differ a lot for some sentences. In the case of sentence 91 this is due to the large unrealistic estimated lag of 1736 ms, meaning that most of the signal is truncated and the small remainder of about 1/6 of the audio signal happens to correlate well with the corresponding small piece of the EEG signal. The difference between corr and xcorr is however not very large for most of the sentences, meaning that corr is probably a meaningful correlation estimate.

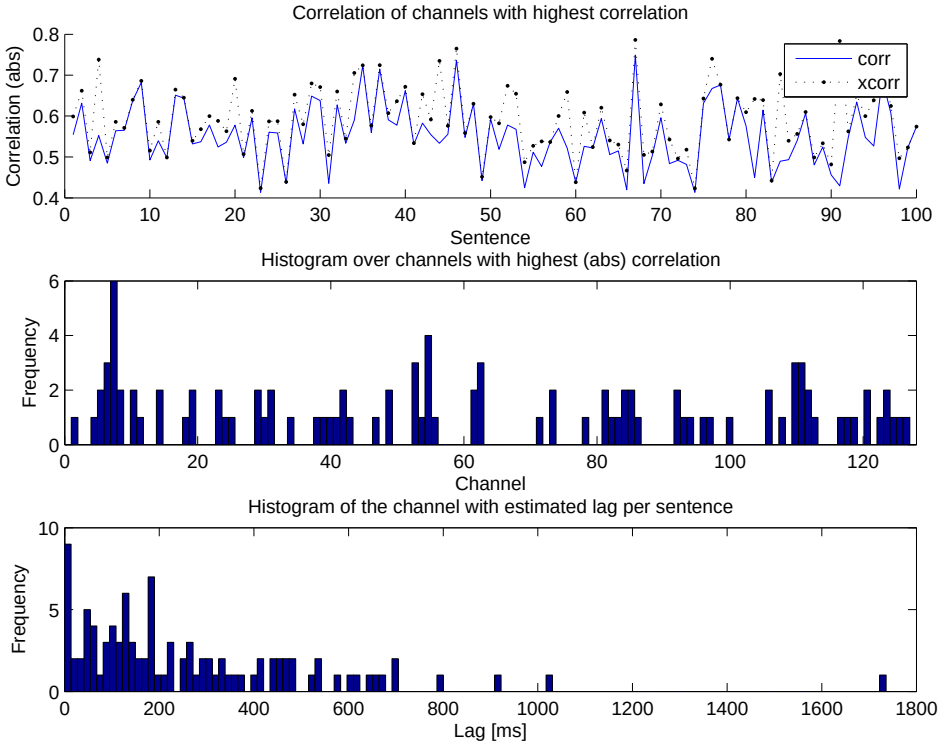


Figure 3.3: *The top plot shows the two different correlation measurements for each sentence. The middle plot illustrate which channels that seem to be best for each sentence. The bottom plot illustrates the distribution of the estimated lags across all sentences. (person 1)*

The concluding remark is that the highest correlating channels correlate poorly, even when compensating with each sentences' estimated lag. This approach is not near good enough for an adequate estimate of the sound envelope. Models have to be used when working in the time domain. Perhaps an approach assuming non-stationarity will yield better results.

3.2 Coherence analysis

Coherence measures the correlation between each frequency component of two signals. If the system is linear, the coherence is the power transfer function between the two signals. Additionally if the two signals are ergodic as well, then coherence can be used to estimate the causality of the signals. Apart from providing another viewpoint of correlation compared with section 3.1, by working in the frequency domain instead of in the time domain, it might provide clues as to frequency bands of additional interest.

Stationarity is assumed here, which may result in underestimating (or overestimating) the values as the formulations might be inappropriate. There exists extensions to coherence that handles time-varying spectral variations of non-stationary signals, see e.g. White & Boashash (1990). In Zhan et. al. (2006) the authors applies such extensions on EEG data.

The coherence between two zero mean signals (x, y) is estimated by

$$\mathsf{coh}_{xy}[f]) = \frac{|\mathsf{P}_{xy}[f]|^2}{\sqrt{\mathsf{P}_{xx}[f]\mathsf{P}_{yy}[f]}}, \tag{3.2}$$

where P_{xy} denotes the cross-spectral density between x and y, and P_{xx} and $\mathsf{P}yy$ denote the auto-spectral density of x and y, respectively. The cross-spectral density is estimated by

$$\mathsf{P}_{xy}[f] = \sum_{n=0}^{N-1} \mathsf{C}_{xy}[n] e^{-i2\pi f \frac{n}{N}}, \tag{3.3}$$

where C_{xy} denotes the cross-covariance of the signals x and y.

The results from the coherence analysis was poor. The averaged estimate of the coherence over a single persons sentences is almost flat and well below 0.2, with an even flatter result if the average is taken over simultaneously all sentences and all channels of a single person. Despite this, the average over all channels for single sentences showed some interesting regions having up to 0.4 in coherence. This is too low for a meaningful reconstruction, but it does show that different sentences contain different important frequency components. No deeper study regarding this was done, but it could possibly yield some more insight to cluster sentences which are similar in this respect. There probably are some kind of separation between the sentences spoken by a male to those spoken by a female. Studying such separations and recalculating the coherence on each separately would be a natural continuation.

3.3 Canonical Correlation analysis

Canonical correlation analysis finds two sets of basis vectors, one for each signal, that maximize the correlation of the two signals when projected onto respectively basis vectors. As a result of the maximization of the projections, canonical correlation is invariant to affine transformations of the signals. This is an outstanding property compared to ordinary correlation analysis in which the correlation is highly dependent on the basis in which the signals are described (Borga, 1998).

The result of CCA is a canonical base $\{w_x, w_y\}$ from the zero mean signals $\{x, y\}$ such that the canonical correlation $p = \mathrm{corr}(w_x x,\, w_y y)$ is maximized,

$$p = \frac{E[w_x^T x y^T w_y]}{\sqrt{E[w_x^T x x^T w_x] E[w_y^T y y^T w_y]}} = \frac{w_x^T C_{xy} w_y}{\sqrt{w_x^T C_{xx} w_x^T w_y^T C_{yy} w_y}}, \tag{3.4}$$

where C_{xx}, C_{yy}, C_{xy} denote covariance matrices of x and y. The maximum of p with respect to w_x and w_y is the maximum canonical correlation and it is found by solving the eigenvalue equations,

$$\begin{cases} C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{yx} w_x = p^2 w_x \\ C_{yy}^{-1} C_{yx} C_{xx}^{-1} C_{xy} w_y = p^2 w_y \end{cases}, \tag{3.5}$$

where w_x denotes the eigenvector corresponding to the largest eigenvalue p^2 of $C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{yx}$ and analogously for the second equation. p^2 as well as w_x and w_y are determined in a straight forward fashion using Singular Value Decomposition (Nordberg, 2012).

In our case, the basis of the audio envelope signal is a scalar and the basis of the EEG signal is a vector with as many dimensions as channels. The later basis vector contains the scalar weights of each channel so that when multiplied with respective channel and summed we get the linear combination of channels that maximally correlate with the audio envelope.

The audio envelope and EEG are BP filtered similarly as before, but here between 2 Hz and 10 Hz due to the amount of noise in the data above 10 Hz. We mainly focus on the 30 selected channels. When using all 128 channels the canonical correlation coefficients reached the order of 10^{10}, a possible sign of over fitting.

To find the highest correlations within a reasonable lag range was CCA calculated for each sentence at lags in the range of 0 to 400 ms with a resolution of 4 ms. It is performed by shifting the two signals 4 ms at a time and in effect making the overlapping signals 4 ms smaller, identically to how the lag-dependent correlation is calculated in section 3.1. Adjusting each sentence for its estimated lag show noticeable improvements of the canonical correlation, as seen in figure 3.4. A substantial part of the sentences' estimated lags are rather evenly distributed, so only a barely noticeable improvement was seen by adjusting all sentences by a globally estimated lag.

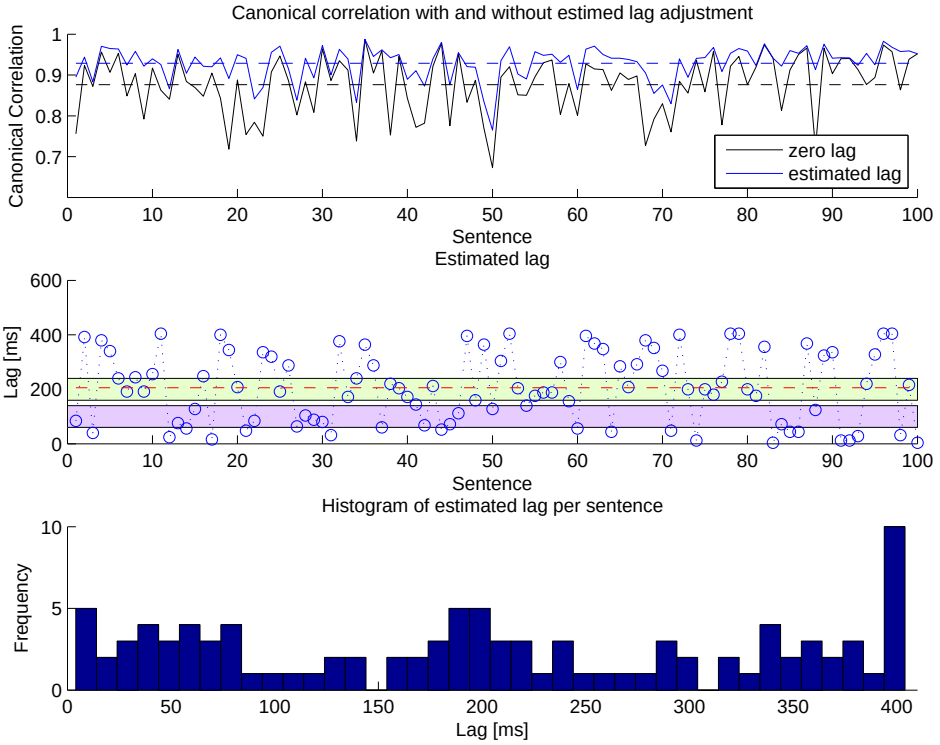


Figure 3.4: *A huge improvement is seen in the top plot when adjusting all sentences to their individually estimated lag. In the middle and lower plot we see the diversity of the estimated lags. In the middle plot the lower lag interval is the N100 region and the upper is the P200 region. Mean values are shown as dashed lines.*

The mean canonical correlation, using the 30 selected channels, is 0.87 without adjusting for individual sentences estimated lag and 0.93 when doing the adjustment. As a comparison, channel 88 was included in exchange for the least significant channel among the 30 selected. The mean canonical correlation then became 0.88 without lag adjustment and 0.93 with lag adjustment. The correlation values are rather similar, indicating that channel 88 doesn't contribute any new significant accessible information which isn't already available though the other channels.

The mean estimated lag over all sentences, for each person, has a very high variance making it hard to draw any conclusions. The canonical correlation depending on lag is shown in figure 3.5. N100 and P200 are clear peaks in general audio-stimulated ERPs and it may therefore be expected to see some correlation in these regions. Small indications of such correlations can be seen in the figure, mostly in P200 for person 1, but they are very small by large and with high un-

certainty. The ERP-range can vary between persons and might explain some of the differences seen.

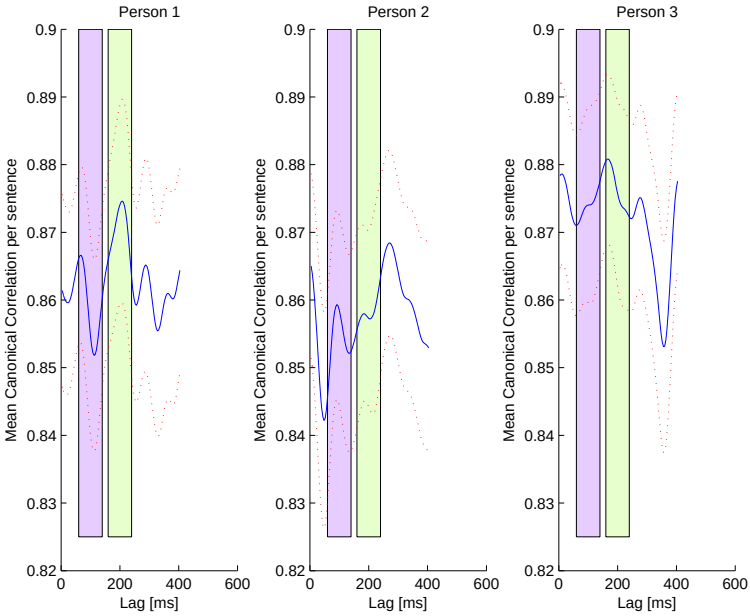
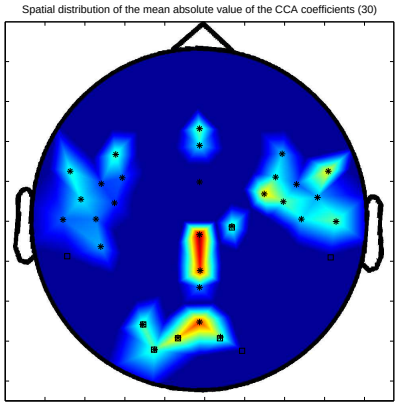
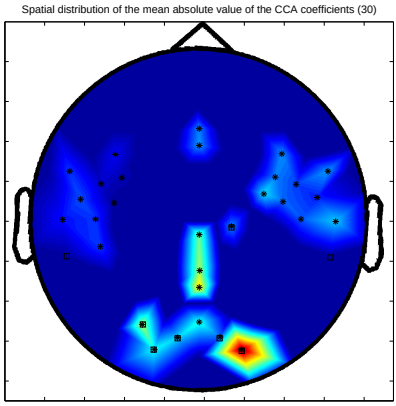


Figure 3.5: *The lower lag interval is N100 and the above is P200. The dotted lines are 99% confidence intervals. (Person 1).*

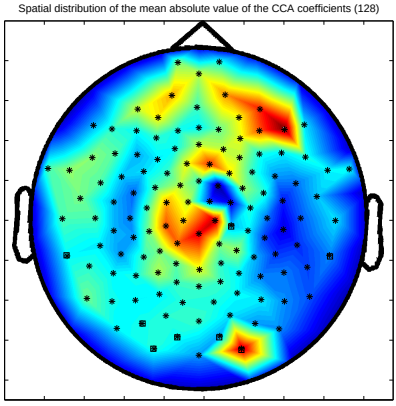
Although the impedance of channel 88 made it less likely to be significant, it was later shown that it actually is considered very significant to CCA for person 1. Figure 3.6 shows a topological plot of the importance of different coefficients when considering all sentences of person 1. When the least significant channel in the 30 selected is exchanged for channel 88, it can be seen that channel 88 dominates completely. It seems that channel 88 contains information which makes some of the information in the other in-ear electrodes, as well as those placed over auditory cortex, redundant, significantly lowering the other channels contributions in comparison. Since this is the case for the electrodes from both sides of the head, it is at least not obvious that the common information is from a common local noise source. The correlation between in-ear electrodes on opposite side do not deviate from the mean correlation between pairs of mirrored electrodes on the scalp, reducing the possibility of some global noise sources such as picking up the acoustic sound waves. Channel 88 is considered very important in the 128 channel case as well. For the other subjects there was no such dominance, although the in-ear electrodes is seen to contribute significantly over most other of the 30 selected channels. Common for all subjects is that a region between the nose and the right ear is significant when using all 128 channels. For one of the subjects the in-ear electrodes contribute far less when using all 128 channels.



(a) *30 selected channels*



(b) *30 selected channels with channel 88*



(c) *128 channels*

Figure 3.6: *The spatial distribution of the mean absolute value of the canonical coefficients. (Person 1).*

Calculation of CCA over all sentences simultaneously was also explored. This was done in a fashion of stacking the sentences, letting the observation of the input and output at each time t be represented by one observation from each sentence's input and output. Calculating CCA over this stack however produced very poor results, even if each sentence was adjusted with its estimated lag. Some other formulation instead of the naive stacking might be better due to the possibility of non-stationary behavior. Other reasons for the poor results could be that the data may still contain too much noise and too many outliers for CCA to have a chance when applied to multiple sentences. There exists more advanced models building on CCA, such as Viinikanoja, Klami & Kaski (2010), which are much more robust to noisy data and the appearance of outliers.

4

L1-regularized linear spectrogram model

We would like to transform the EEG multi-channel signal into an estimate of the single channel envelope of the heard audio. A spectrogram describe the time variate frequency energy content of a signal. With speech being frequency energy distributions interchanged in time, this make such a signal basis interesting to investigate. Spectrogram approaches have been used by others in similar settings with good results, see Mesgarani & Chang, 2012. Furthermore, it is desired that such a model do not use all channels, as not all of them should be enough related to the sound, and a lot of redundancy exists due to spatial closeness of the sensors. By the use of a L1-regularization term the number of channels can be controlled and forced to a minimum with respect to the accuracy of the estimations produced by the model. The regularization also counter over-fitting issues if adjusted sufficiently.

To estimate the spectrogram it is necessary to describe the frequency content over time. This is done by the discrete time Short-time Fourier transform (STFT), equivalent to the Discrete Time Fourier Transform (DTFT) truncated by a moving window, which is estimated by

$$\mathsf{STFT}_{x[n],w[n]}(m,f) = X(m,f) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-i2\pi nf}, \tag{4.1}$$

where $x[n]$ denotes the input, $w[n]$ denotes a window function (usually Hamming). Here m and f denote time and frequency components, respectively.
The spectrogram of a signal, given a window function $w[n]$, is then estimated as the magnitude squared of the STFT of that signal,

$$\mathsf{Spectrogram}_{x[n],w[n]}(m,f) = |X(m,f)|^2. \tag{4.2}$$

Least Squares is used to solve for the linear model. Let x, y denote known column vectors and w an unknown matrix such that

$$y = w^T x + e, \tag{4.3}$$

where e denotes Gaussian distributed noise. Suppose that $\hat{w}$ is an estimate of w, the residuals of the projection of x on w is given by

$$\text{Residuals} = (y - \hat{w}^T x)^T (y - \hat{w}^T x) = (y - \hat{w}^T x)^2 = (y - \hat{y})^2. \tag{4.4}$$

The residuals will tend to zero as $\hat{w}$ models more and more of e. This is called overfitting and it will lead to very bad generalization for new data. It may be the case that w is insufficient to fully model e and so the residuals will never be zero. The least squares solution

$$\hat{w} = \underset{w}{\operatorname{argmin}} \left(y - w^T x\right)^2, \tag{4.5}$$

is the $\hat{w}$ that minimizes the residuals in the presence of Gaussian noise.
To handle overfitting a regularization term is often introduced which depend on $\hat{w}$ in some fashion. A regularization is generally given by

$$\hat{w} = \underset{w}{\operatorname{argmin}} \left(y - w^T x\right)^2 + \mathcal{J}(\lambda, w). \tag{4.6}$$

There exists a variety of regularization terms with different purposes since it is a way to introduce additional information, a priori, into the optimization formulation. Our a priori is that we assume a sparse solution, meaning that we presume that far from all EEG channels are actually needed to estimate the sound envelope of heard speech. This is assumed due to the anatomical knowledge of the brain, where the audio cortex resides and from previous studies. We therefore make use of regularization terms which encourage sparsity in $\hat{w}$, rather than a residuals minimizing $\hat{w}$ which could use more or less all EEG channels.

Least Absolute Shrinkage and Selection Operator, LASSO, is a L1-norm regularization of least squares (Bishop, 2006), (Hastie, Tibshirani & Friedman, 2008). The regularization term is the L1-norm of the weight matrix w,

$$\mathcal{J}(\lambda, w) = \lambda \|w\|_1, \tag{4.7}$$

and it forces w to become sparse, depending on the regularization parameter λ as a user parameter. The general LASSO regularized Least Squares formulation is then

$$\hat{w} = \underset{w}{\operatorname{argmin}} \left(y - w^T x\right)^2 + \lambda \|w\|_1. \tag{4.8}$$

4.1 Model

Let $\mathcal{S} = \{1, 2, ..., s\}$ be a set of sentences with corresponding EEG signals and $\mathcal{C} = \{1, 2, ..., c\}$ be a set of EEG channels of an EEG signal.

The general form of linear combinations of spectrograms is

$$|Y[m, f]|^2 = \sum_{c \in \mathcal{C}} \beta_c[m, f]|X_c[m, f]|^2, \tag{4.9}$$

where m and f denote time and frequency indexes, respectively. Here β denotes the unknown vector which consist of a scalar weight for each channel's spectrogram. The weight is frequency and time dependent, meaning that the linear combination of the channels' spectrogram changes over both time and frequency.

The application of the more specific case of time and frequency independent linear combinations of the spectrogram,

$$|Y[m, f]|^2 = \sum_{c \in \mathcal{C}} \beta_c|X_c[m, f]|^2, \tag{4.10}$$

is studied in this report. This can be posed as a least squares problem

$$\underset{\beta}{\operatorname{argmin}} \sum_{s \in \mathcal{S}} \left(|Y_s|^2 - \sum_{c \in \mathcal{C}} \beta_c|X_{s,c}|^2\right)^2. \tag{4.11}$$

By introducing a LASSO regularization over the channels of β, the minimization is also optimized over the number of channels used, forcing β_c to go towards zero for some channels. The problem formulation follows as

$$\underset{\beta}{\operatorname{argmin}} \sum_{s \in \mathcal{S}} \left(|Y_s|^2 - \sum_{c \in \mathcal{C}} \beta_c|X_{s,c}|^2\right)^2 + \lambda\|\beta\|_1, \tag{4.12}$$

where λ denotes the regularization parameter. This can be re-formulated as the general LASSO formulation

$$\hat{w} = \underset{w}{\operatorname{argmin}} \left(y - w^T x\right)^2 + \lambda\|w\|_1, \tag{4.13}$$

where w denote a column vector, $w = \begin{bmatrix} w_1 & \dots & w_C \end{bmatrix}^T$, which contain the scalar weighting of the different channels (equivalent to β). x denote a matrix,

$$x = \begin{bmatrix} \begin{bmatrix} \mathsf{spec}^1_{1,1} & \dots & \mathsf{spec}^N_{1,1} \end{bmatrix} & \dots & \begin{bmatrix} \mathsf{spec}^1_{1,S} & \dots & \mathsf{spec}^N_{1,S} \end{bmatrix} \\ \vdots & \ddots & \vdots \\ \begin{bmatrix} \mathsf{spec}^1_{C,1} & \dots & \mathsf{spec}^N_{C,1} \end{bmatrix} & \dots & \begin{bmatrix} \mathsf{spec}^1_{C,S} & \dots & \mathsf{spec}^N_{C,S} \end{bmatrix} \end{bmatrix}, \tag{4.14}$$

with a outer column for each sentence s and a row for each channel c, with each such column-row-cell being a row-vector $\mathsf{spec}_{c,s}$ of a vectorized spectrogram for that channel. y denote a row vector,

$y = \left[\left[\mathsf{spec}_1^1 \quad \dots \quad \mathsf{spec}_1^N\right] \quad \dots \quad \left[\mathsf{spec}_S^1 \quad \dots \quad \mathsf{spec}_S^N\right]\right]$, with each outer column as a row vector spec_s of a vectorized spectrogram of the audio envelope of sentence s.

4.2 Single sentence

The sound envelope of a sentence is calculated as described in chapter 3.1. The spectrograms of the sound envelope and of all the EEG channels are calculated using a sliding Hamming window of length $\lceil \frac{10}{250} \rceil \times N$, where N denotes the signal length. The length is chosen to be small enough to precisely capture frequencies of 10 Hz and below. The EEG signal is sampled at 250 Hz which the audio envelope is down sampled to as well. The signals then have the same length. A model is estimated for each sentence consisting of a weight vector β with a scalar value for each channel of the EEG signal. Residuals and explained variance are used to compare the estimated spectrograms with the original, see appendix A for details on these statistics.

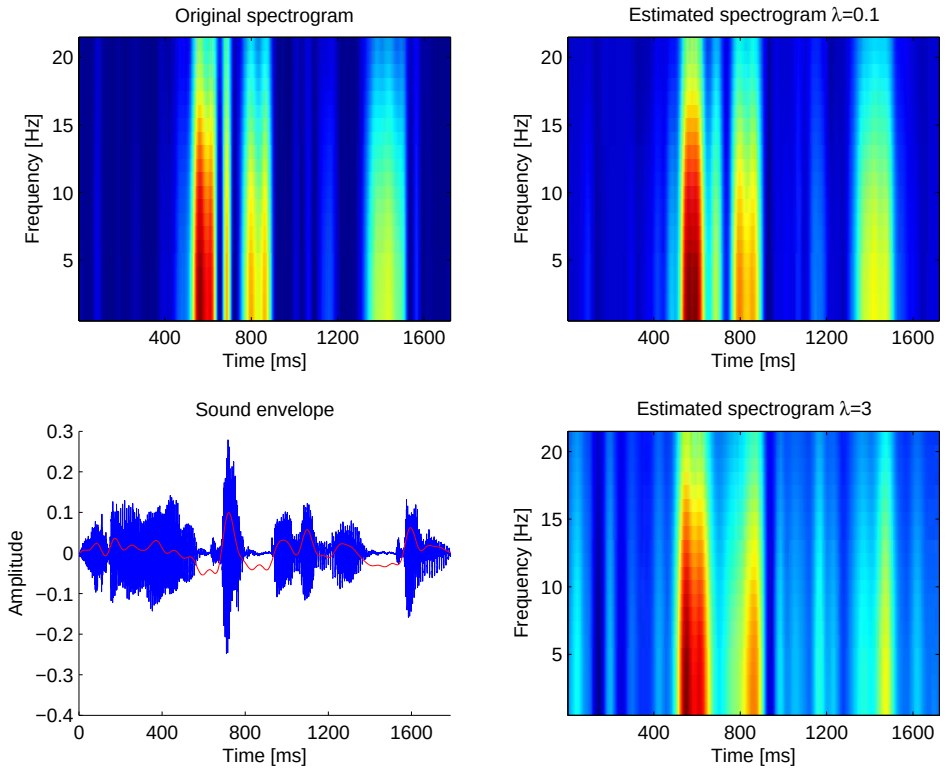


Figure 4.1: *A sound envelope and its spectrogram is to the left. To the right are spectrogram estimates from the corresponding EEG, for two different λ values. (Person 1, sentence 4.)*

The result of applying this method to a single sentence is seen in figure 4.1. The effect the regularization parameter have on how well we can model the original spectrum is shown for two such values. The number of channels necessary for

each of the two models are shown in figure 4.2. The model estimated using regularization parameter value 0.1 is quite similar to the original spectrogram, with an explained variance of 0.98, and with 0.00 as residuals. The residuals are small enough to make us cautious of possible overfitting, while only requiring 44 non-zero channels. The model estimated using the regularization parameter value 3 is clearly worse at modeling the original spectrogram, while the most distinctive features are still seen present. The explained variance of this model is 0.80, and with 0.02 as residuals, while however only requiring 16 non-zero channels. Figure 4.3 show the head topology of the importance of the different channels.

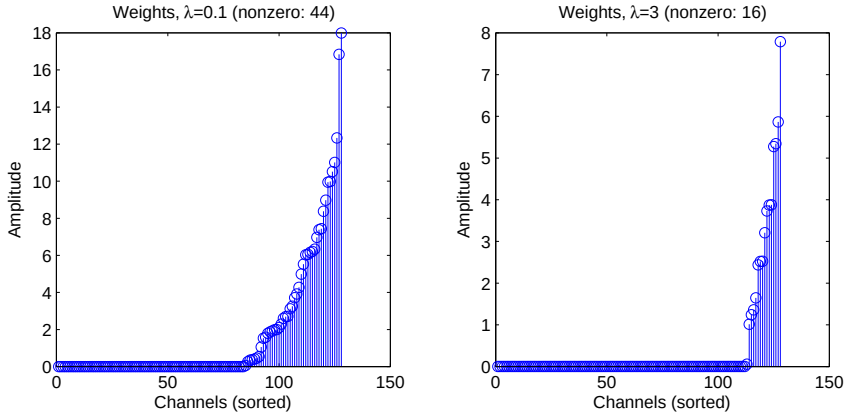


Figure 4.2: *Weights (β_c) in the two models used for the estimation in figure 4.1, sorted on size. A channel c is considered to be significant (non-zero) if $|\beta_c| >$* 0.0001.

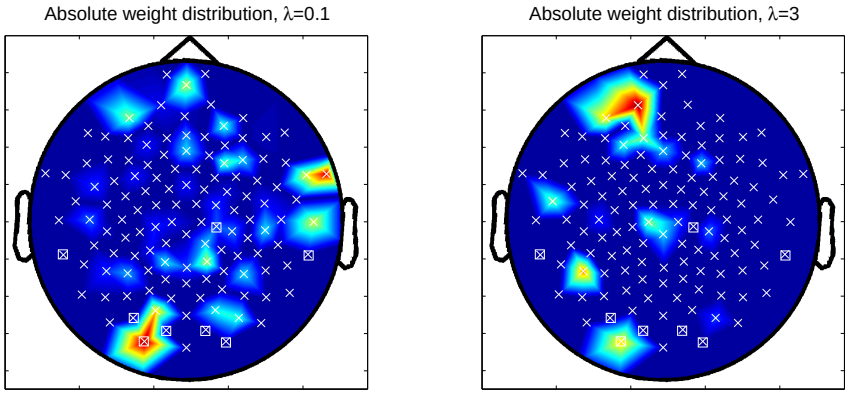


Figure 4.3: *Spatial distribution of the absolute values of the weights in figure 4.2.*

An important aspect is to determine how many EEG channels that are needed to get a useful estimation of the sound envelope. The number of channels that are non-zero is affected by the value of the regularization parameter λ. In figure 4.4 the regularization parameter is plotted against the explained variance of the model and the number of non-zero channels ($\beta_c > 0.0001$) respectively, for all sentences of person 1. The number of non-zero channels drops significantly faster than the explained variance with a increasing value of the regularization parameter. For values of $\lambda > 5$ the curves plane out into a slow linear decrease of explained variance and non-zero channels. Figure 4.5 shows head topologies of the placement of the mean of the most important channels over all sentences of person 1.

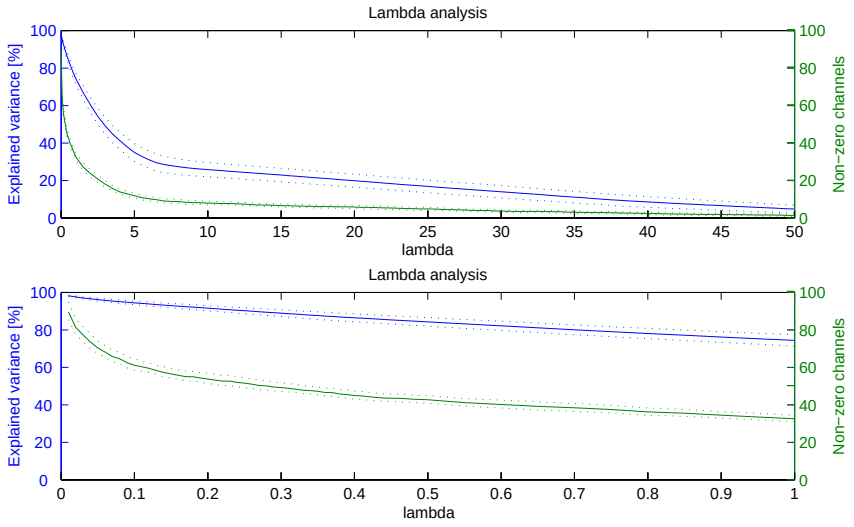


Figure 4.4: *Explained variance and the number of none-zero channels as a function of the regularization parameter λ. 99% confidence interval indicated by dotted lines. (Person 1.)*

The colors in the spectrograms in figure 4.1 highlight temporal-stretched local frequency energy minimums. These seem to be robust and might be, if enough diversifying, an appropriate feature to use when distinguishing between different sentences. This has however not been verified in this study.

In figure 4.3 it can be seen that channel 73, placed in the ear, is of great importance to the shown model instances. In figure 4.5 we see that also in the mean importance of the channels some channels placed in the ears are very important. This is especially true for channel 88 in the regularization parameter value range 0.5 to 6, which seems to be in the range of reasonable values in the sense of avoiding over-fitting.

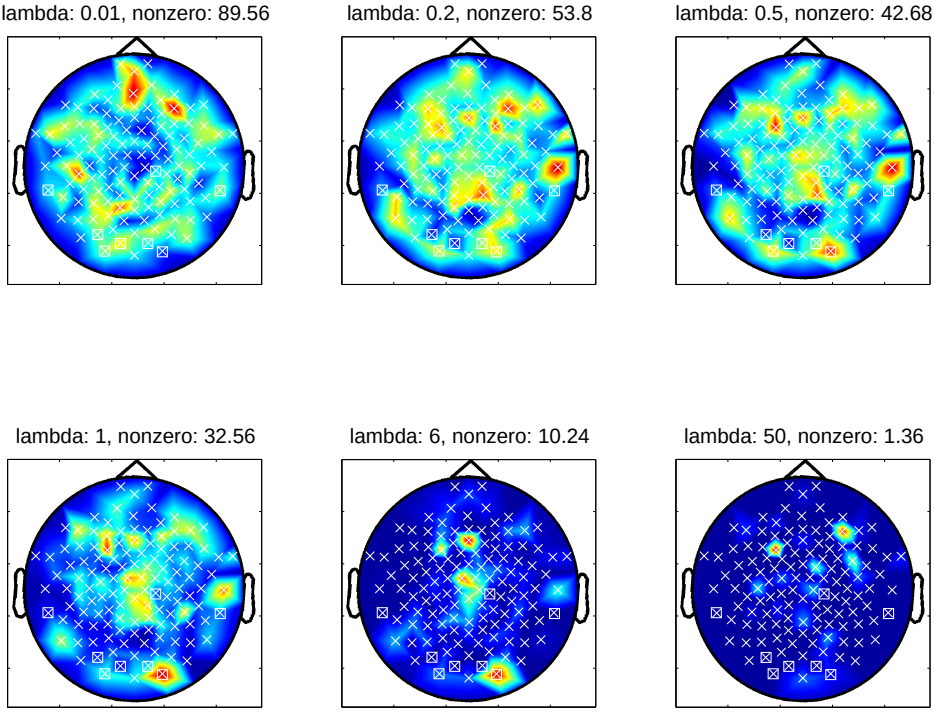


Figure 4.5: *Mean over all sentences of the absolute value of the respective coefficients. The different images show different number of (mean) non-zero channels due to different λ values (higher λ give fewer channels used). Note the island at the bottom right in the fifth plot, it is channel 88 located in the ear.*

4.3 Multiple sentences

Solving for β for multiple sentences was attempted by simply stacking the data from the different sentences' on top of each other, without any modifications. This is illustrated in figure 4.6 using $\lambda = 1.0$, which is a reasonable value maintaining high explained variance using few channels, as seen in figure 4.4. Whereas the model have no problem to explain a single sentence using fewer than 20 channels, it quickly breaks down as early as with 10 simultaneously explained sentences, using about 70 channels for 10 or more sentences. This show possible limitations in the model which could possibly be addressed by using a more dynamic model.

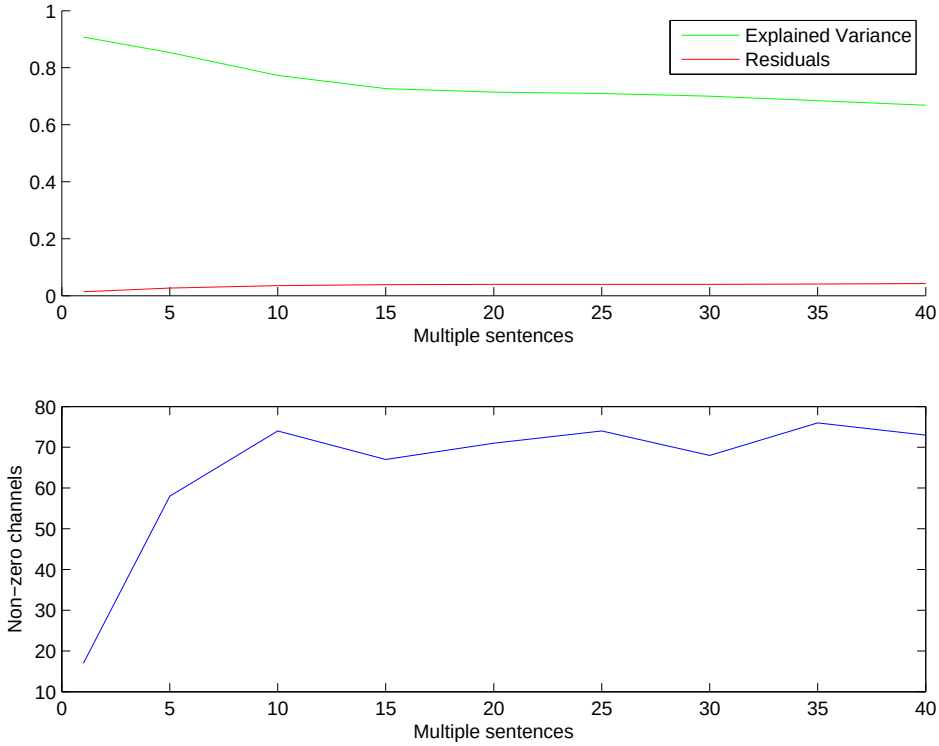


Figure 4.6: *Shows how well the model can explain multiple sentences simultaneously for different number of sentences. Using $\lambda = 1$. The n first sentences from the trial of person 1 were used where $n \in \{1, 5, 10, 15, 20, 25, 30, 35, 40\}$*

Using this approach and our implementation we could not try higher sentences amount due to limitations in RAM, with 16 GB not being enough.

The mean importance of the channels is shown in figure 4.7. Some minor resemblance to 4.6 can be seen, especially the importance of channel 88. The time and

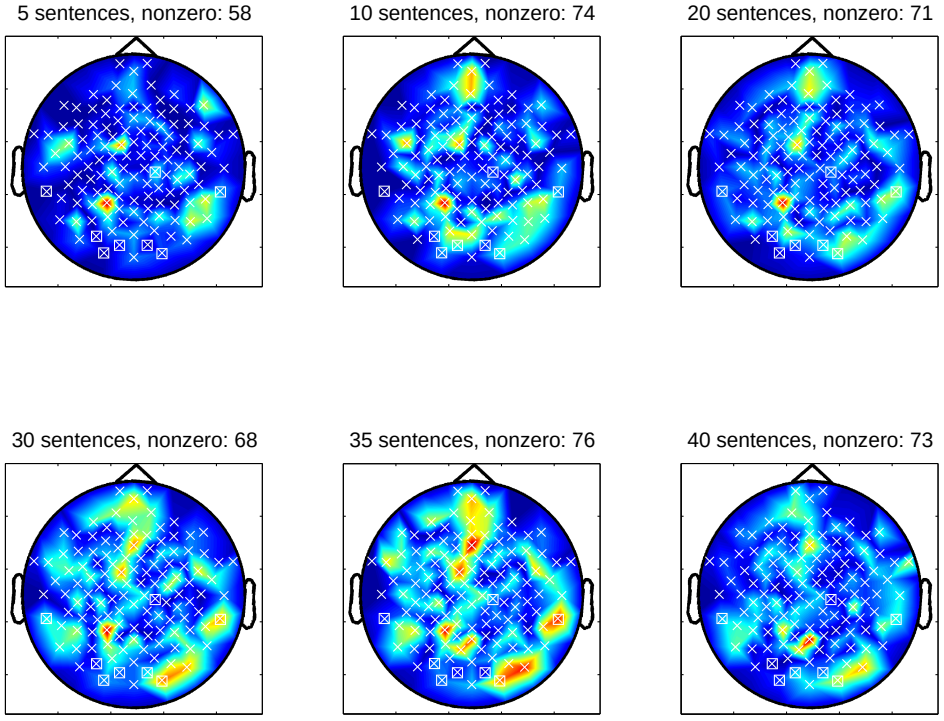


Figure 4.7: *Head topology corresponding to Figure 4.6. The absolute value of the coefficients when using the first* $n \in \{5, 10, 20, 30, 35, 40\}$ *sentences to train on.*

frequency independent linear combinations of the spectrogram is sufficient to explain single sentences but does not generalize well. It is not implausible that the time and/or frequency varying models might provide further insight and better generalization behavior. This since they can model more of the possibly non-stationarity behavior such as temporal transients. Previous works have used models involving using frequencies varying in time with good results (Mesgarani & Chang, 2012).

5

L1-regularized linear filter model

Power A. J., et. al. (2012) overcomes previous techniques' limitations in real time applicability by utilizing auditory evoked spread spectrum analysis (AESPA) and manages to show attentional effects of attending two continuous natural speech streams. AESPA is a linear filter in the time domain with the EEG as input and the logarithm of the sound envelope as output, which they estimate using least squares. Lalor et. al. (2012) archives very high accuracy of auditory attention classification using AESPA single channel inversion and even higher using all channel inversion. Inspired by these results we estimate a linear filter using least squares and regularized using the group L1-norm to minimize the number of channels used. The data set used in the two named studies consists of 60 second trials while our data is limited to about 2.6 second trials. This may limit the ability of the models to generalize to the data but still allows for evaluation of how many channels that are needed while explaining the data sufficiently. The statistics used are defined in Appendix A.

Regularized Least Squares is used to solve for the linear model. Regularized Least Squares is introduced in section 4 as

$$\hat{w} = \underset{w}{\operatorname{argmin}} \left(y - w^T x\right)^2 + \mathcal{J}(\lambda, w), \tag{5.1}$$

together with the LASSO regularization term

$$\mathcal{J}(\lambda, w) = \lambda \|w\|_1. \tag{5.2}$$

In this section a generalization of LASSO is used. This generalization is known as Group LASSO and it is LASSO performed over clusters of coefficients, thereby making w sparse in the number of such clusters (Meier, Geer & Bühlmann, 2008).

This is archived by summing over the L2-norm of each cluster,

$$\mathcal{J}(\lambda, w) = \lambda \sum_c \|w_c\|_2. \tag{5.3}$$

The general Group LASSO regularized Least Squares formulation is then

$$\hat{w} = \underset{w}{\operatorname{argmin}} \left(y - w^T x\right)^2 + \lambda \sum_c \|w_c\|_2. \tag{5.4}$$

5.1 Model

Let $\mathcal{S} = \{1, 2, ..., s\}$ denote a set of sentences with corresponding EEG signals and $\mathcal{C} = \{1, 2, ..., c\}$ denote a set of EEG channels of an EEG signal.

The general form of a linear filter applied to a signal in the time domain is

$$y(t) = (w * x)(t) + e(t), \tag{5.5}$$

where $y(t)$ denotes the sound envelope which is equal to $e(t)$ added with the convolution between $w(\tau)$ and $x(t)$. Here, $x(t)$ denotes the EEG signal, $w(\tau)$ denotes a linear filter and $e(t)$ is Gaussian distributed white noise. The problem of estimating w can be posed as a least square problem

$$\underset{w}{\operatorname{argmin}} \sum_{s \in \mathcal{S}} \left(y_s(t) - \sum_{c \in \mathcal{C}} (w_c * x_{s,c})(t)\right)^2. \tag{5.6}$$

By introducing a group LASSO regularization over the channels of w the minimization is also optimized over the number of channels used, forcing w_c to be zero for some channels. The problem formulation follows as

$$\underset{w}{\operatorname{argmin}} \sum_{s \in \mathcal{S}} \left(y_s(t) - \sum_{c \in \mathcal{C}} (w_c * x_{s,c})(t)\right)^2 + \lambda \sum_{c \in \mathcal{C}} \|w_c\|_2, \tag{5.7}$$

where λ denotes the regularization parameter. When the length of the filter w is large it is much more efficient to solve the problem in the frequency domain by re-formulating it as

$$\underset{W}{\operatorname{argmin}} \sum_{s \in \mathcal{S}} \left(Y_s(t) - \sum_{c \in \mathcal{C}} W_c X_{s,c}\right)^2 + \lambda \sum_{c \in \mathcal{C}} \|W_c\|_2, \tag{5.8}$$

where Y, X and W denote the respective Fourier transforms of y, x and W. Since the regularization still acts like a L^1-regularization over the L^2-norm of each channel, the minimization is still optimized over the number of channels used just like before.

This filter can be approximated as a Finite Impulse Response (FIR) filter with a certain length N, or as a discrete transfer function of length N of the problem formulation in the Fourier domain. There is therefore two design parameters, the regularization parameter λ and the filter length N.

For solvers requiring the unknown variable to be a vector, this can be re-formulated as the general Group LASSO formulation

$$\hat{W} = \underset{W}{\operatorname{argmin}}(Y - W^T X)^2 + \lambda \sum_{c \in \mathcal{C}} \|W_c\|_2, \tag{5.9}$$

where $Y = \begin{bmatrix} y_1 & \dots & y_S \end{bmatrix}^T$ is a column vector containing all the sentence's envelope's Fourier transforms y_s appended after each other. $W = \begin{bmatrix} w_1 & \dots & w_C \end{bmatrix}^T$ is a column vector containing all the filters of each separate channel appended after each other. The length N of the filter w_c, which is the same length for each channel, is a design parameter. X denote a matrix,

$$X = \begin{bmatrix} \begin{bmatrix} X_{1,1}^1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & X_{1,1}^N \end{bmatrix} & \dots & \begin{bmatrix} X_{1,S}^1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & X_{1,S}^N \end{bmatrix} \\ \vdots & \ddots & \vdots \\ \begin{bmatrix} X_{C,1}^1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & X_{C,1}^N \end{bmatrix} & \dots & \begin{bmatrix} X_{C,S}^1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & X_{C,S}^N \end{bmatrix} \end{bmatrix}, \tag{5.10}$$

containing matrices with Fourier transforms of the EEG-signal of each channel on the diagonal, with such channel-matrices as rows for each channel, and all sentences are appended as such columns. The Fourier transforms are reduced of higher frequency content to match the filter length N. The solving is made many times more efficient in the frequency domain compared to in the time domain. The formulation also allow for the use of sparse matrices which makes it even more efficient.

5.2 Single sentence

The $\log_{10}$ of the sound envelope is calculated as described in chapter 3.1. Single sentences was first used to train a model for each sentence, in order to see if it can model the individual signals well and to analyze changes in regularization parameter and filter length. The model is much more expressive than any other tested in this study since it is both time and context dependent. This makes it very prone to over fitting when modeling any single sentence, even using a very large regularization parameter value as seen in figure 5.1 or a very short filter length as seen in figure 5.2. Very few channels are needed and for some sentences a single channel is enough to explain it accurate using this model. A channel is considered non-zero if its energy is larger than 10^{-8}.

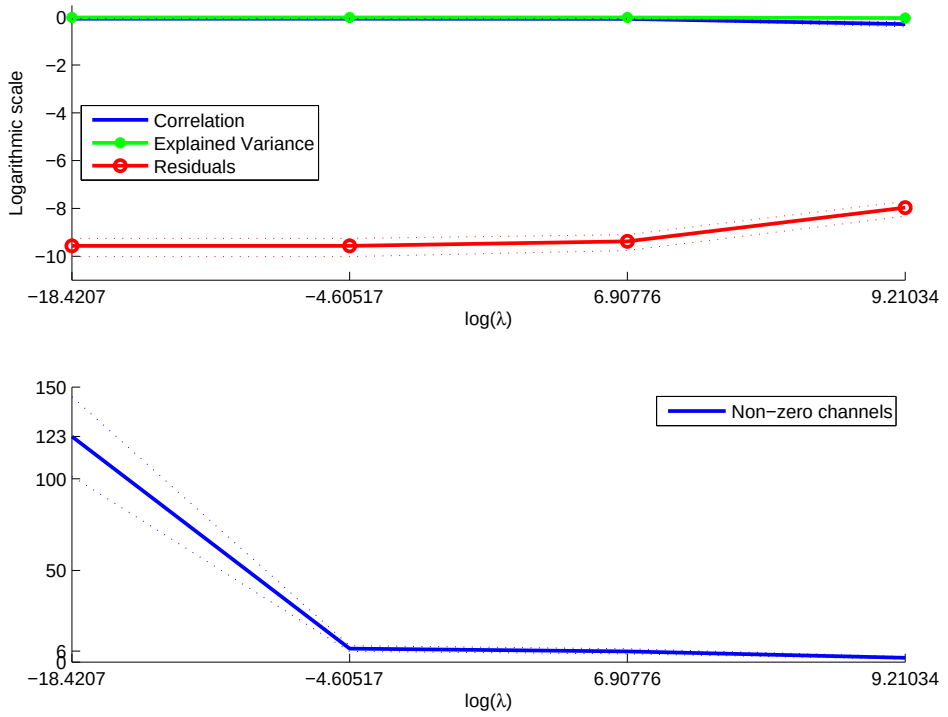


Figure 5.1: *Analysis of the regularization parameter λ using a filter length of 120ms. 99% confidence interval are indicated by dashed lines. (Person 1.)*

We can also see that the model easily capture the variance of the signal, making the Explained Variance remain about the same when varying λ and the filter length respectively. It clearly captures the same amount of variation but if it is the same kind varies as seen when comparing with the correlation. The small residuals in figure 5.1 and figure 5.2 indicate that a severe overfit is likely to occur, even with reasonably high values of the regularization parameter λ. This

is probably due to the limited length of sentence data used while the model is too expressive.

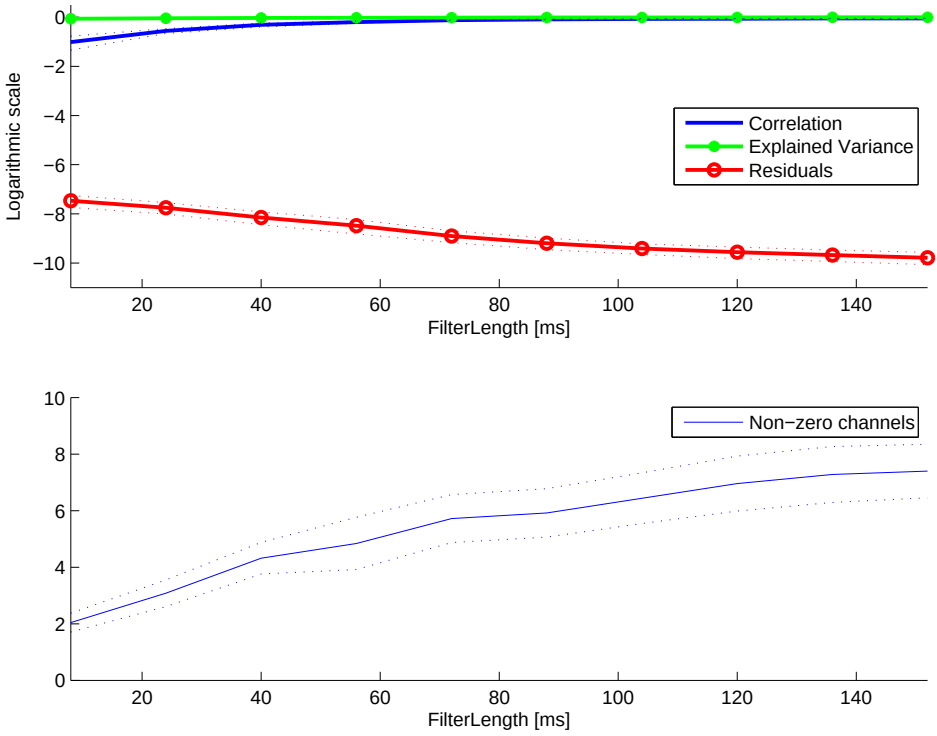


Figure 5.2: *Filter length analysis over the sentences. A 99% confidence interval is indicated by dashed lines. $\lambda = 100$. Since each sample is 4 ms long, the spatial length of the FIR filter is one forth of its millisecond length. (Person 1.)*

The mean importance of the channels is shown in figure 5.3. Remarkably channel 88 is dominating as the most useful regardless of the number of non-zero channels. Channel 113, placed over the right audio cortex, is also deemed significant on average. Compared to the previous model's estimates, the electrodes used are mostly from the out-skirt of the head and especially the neocortex in front.

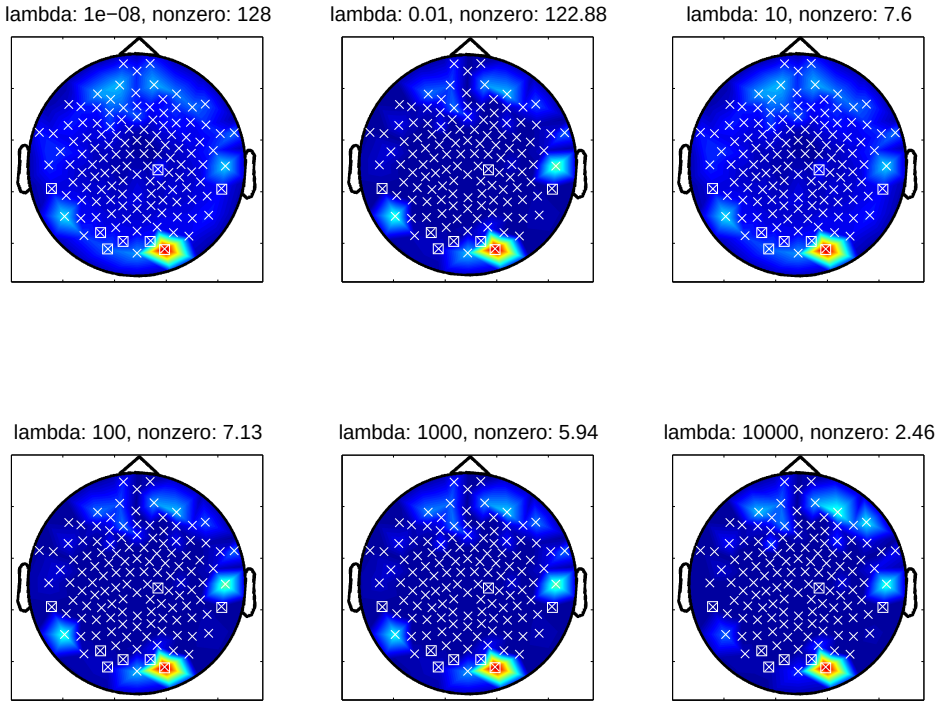


Figure 5.3: *Mean over all sentences of the absolute value of the respective coefficients. The different images show different number of (mean) non-zero channels due to different λ values (higher λ give fewer channels used). Notice the island at the bottom right, it is channel 88 located in the ear.*

5.3 Multiple sentences

Neither does this model generalize well in its current implementation or with the used test data. $\lambda = 100$ enables the model to capture enough to support a correlation of 0.72 for all sentences and keeping steady on 0.92 up till 65 sentences simultaneously. However as with the spectrogram model in section 4 almost all channels are used when modeling many sentences. A higher regularization value of 10000 with fewer non-zero channels is shown in figure 5.4, with corresponding head topology in figure 5.5. The correlation vary from 0.7 to 0.6, with the explained variance kept close to 1.0 and residuals around $0.4 \cdot 10^{-3}$.

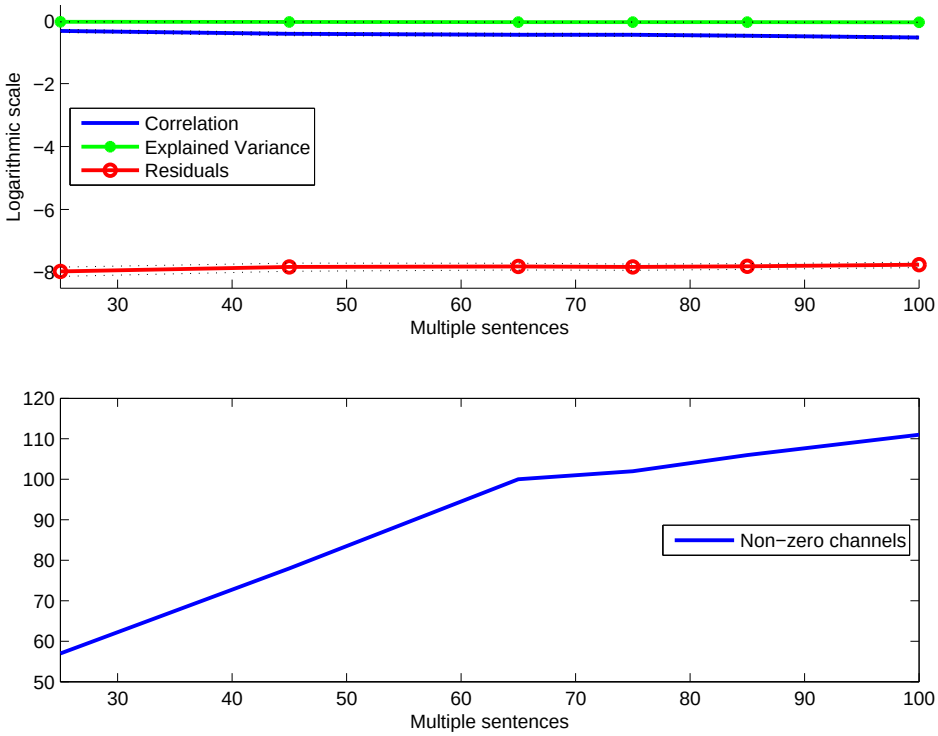


Figure 5.4: *Shows how well the model can explain multiple sentences simultaneously against the number of simultaneous sentences. Using* $\lambda = 10000$ *and a filter length of 120 ms (26 samples long). The n first sentences from the trial of person 1 were used where* $n \in \{25, 35, 45, 65, 75, 85, 100\}$

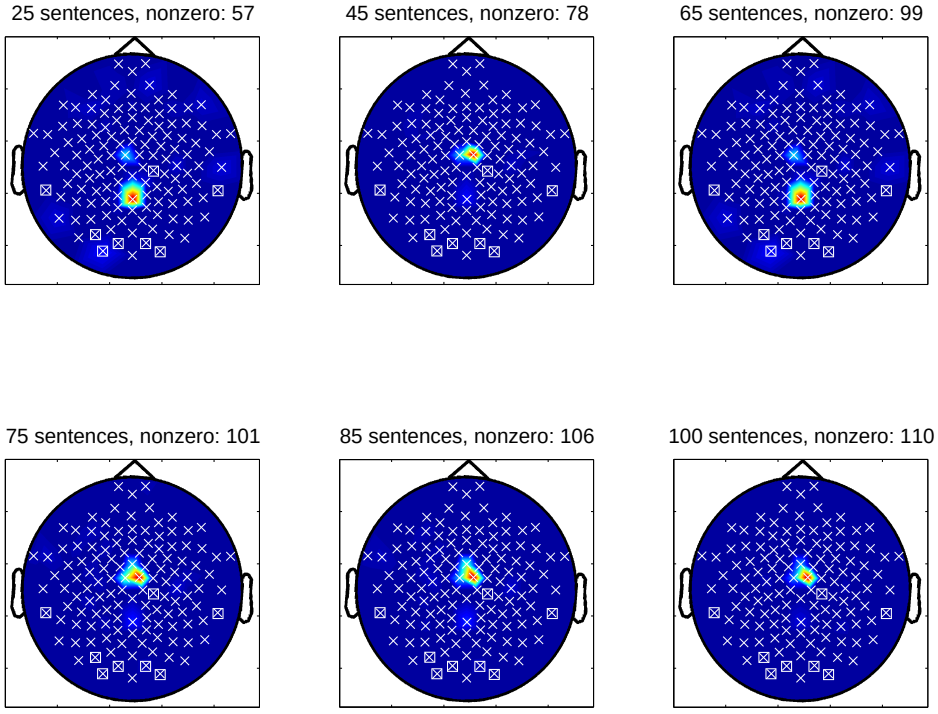


Figure 5.5: *Head topology corresponding to Figure 5.4. The mean energy of each coefficient when using the first* $n \in \{25, 45, 65, 75, 85, 100\}$ *sentences to train on.*

Filters of lengths of over 200 ms (a important delay seen in the works of Power A. J., et. al. (2012)) were explored but they provided little insight. We didn't see anything close to the findings of them nor a clear P200 peak. This is probably due to the persistent overfitting with filters of length 120 ms and due to the limited length of the test data. After down sampling the effective number of discrete values was about 400 per trial, each covering a 4 ms window.

This model could probably performs much better with a lot more data. Further strategies that can be tested are to use auto regression, building on the feedback of the previous calculated output as input. Since no indication of meaningful generalization was seen wasn't more advanced training schemes such as cross validation tried.

6
Conclusions

We have formulated and evaluated two sparse linear models with respect to how well they can explain sound envelopes from EEG. The first model consist of linear combinations of spectrograms. A general form of the model with time and frequency varying weighting of the spectrograms is proposed while a simpler, time and frequency invariant form of the model is trained on the data set and evaluated. The second model is a linear filter model, introduced generally as continuous in time, but implemented as a FIR filter of length N. It is trained in the frequency domain as a discrete transfer function which lowers the computational performance of training significantly, and with efficient use of sparse matrices reducing it even further. Both models are sparse regarding the number of EEG channels used. The sparseness is archived by the introduction of L1-regularization to the Least Squares formulation which forces the models to use as few channels as possible. How few is controlled by the regularization parameter λ, which then relates to how well the model is allowed to describe the training data. This is an automatic channel selection technique, forcing the most redundant and least useful channels' usage to zero.

In table 6.1 the models are compared with the linear Canonical Correlation Analysis (CCA) using the statistics of Correlation, Explained Variance and Residuals (see Appendix A). Neither model managed to generalize to evaluation data not trained on based on the data set in this study. This might be explained by the data sequences of 2.6 seconds on average being to short, or that the models do not contain enough dynamics. All models are evaluated on the training data to explore how well the data can be explained and to analyze the required amount of channels in order to reconstruct the sound envelopes. Both sparse models manage to get rid of much of the redundant information in the EEG electrode grid, significantly reducing the number of used electrodes. The linear filter model

can model every sentence separate to a very high degree of accuracy given only 6.12 channels on average. It does this far better than CCA with all 128 channels as well as with the sentences individually adjusted to lag. The small residuals of the linear filter model do however indicate likely over-fitting problems. The model is in fact very dynamic and would need much more and longer sequences of training data to be rightfully evaluated. This flexibility of the model can be seen in the result of training on multiple sentences. With a filter length of 36 samples (144 ms) it manages to reconstruct the individual sentences to a high degree when using most channels. The less dynamic linear spectrogram model is constant in both time and frequency which makes it far less dynamic than the filter model, but it archives interesting reconstructions using as few channels as 16 for some sentences, as shown in chapter 4, while archiving similar Explained Variance when using 42 channels on average for all sentences individually.

Many avenues for further work has been presented throughout the report. Improved versions of correlation analysis, coherence analysis and CCA have their use motivated and are refereed to in respective section in chapter 3. Although we had no time to evaluate their improvements within our context, they might improve the results as they allow for non-stationary signals, are robuster to noisy data and handle outliers. The coherence analysis performed showed that sentences individually contained important frequency components. Clustering of similar sentences in this respect might provide the possibility of breaking down the problem to multiple models or multi-modal models. The training data could probably be used more efficient if a discriminant model were used (Ng A. & Jordan M., 2001), making use of the fact that there are many negative examples in the form of other sentences. The linear filter used in this study is a Finite Impulse Response (FIR) filter. More advanced linear filters of the class Infinite Impulse Response (IIR) filters might be better at capturing the dynamics in the brain, such as Auto-Regressive (AR) or Auto-Regressive Moving Average (ARMA) filters (Gustafsson, Ljung & Millnert, 2010). ARMA handles easy cases of non-stationarity by adapting to changing averages.

Finally, small positive indications of the usefulness of in-ear EEG electrodes are shown within the scope of this study. Both sparse models as well as CCA make use of the electrodes placed in the ears. For the first subject the ear electrode (channel) 88 is the single most important electrode, on average, among the selected 30 electrodes used by CCA in chapter 3. This electrode is also showing up as being among the most important, on average, in the linear spectrogram model, both in regard to single channels (figure 4.5) and when training over multiple channels (figure 4.7). It is also the most important electrode, on average, in the linear filter model when training on individual sentences (figure 5.3). The other in-ear electrodes are also shown to be contributing, in many cases significantly such as when channel 88 isn't used by CCA. The later implies that they might carry useful but redundant information. A comparison of CCA results with and without channel 88 is shown in figure 3.6. The in-ear electrodes are, on average, the most important or among the most important electrodes according to both sparse models as well as for CCA for person 1. The in-ear electrodes are not as

dominant for the other subjects, while still seen to be important. It is a positive indication of the possibility of using electrodes in the ears for future hearing aids, something that should be further investigated.

Model	Channels	Correlation	Explained Variance	Residuals
CCA, *single sentences*	30	0.87	0.54	$7.44 \cdot 10^{-2}$
single sentences	128	0.82	0.53	$7.49 \cdot 10^{-2}$
CCA - lag adjusted, *single sentences*	30	0.93	0.71	$5.20 \cdot 10^{-2}$
single sentences	128	0.92	0.71	$5.26 \cdot 10^{-2}$
Linear spectrogram, *single sentences*	42	-	0.84	$4.90 \cdot 10^{-3}$
single sentences	32	-	0.74	$8.10 \cdot 10^{-3}$
single sentences	20	-	0.54	$1.49 \cdot 10^{-2}$
single sentences	16	-	0.45	$1.82 \cdot 10^{-2}$
Linear spectrogram, *multiple sentences (40)*	128	-	0.67	$4.29 \cdot 10^{-2}$
multiple sentences (40)	34	-	0.64	$4.73 \cdot 10^{-2}$
multiple sentences (40)	24	-	0.63	$4.86 \cdot 10^{-2}$
Linear FIR filter, *single sentences*	6.12	0.93	0.99	8.460e-05
single sentences	2.52	0.75	0.96	3.470e-04
Linear FIR filter, *multiple sentences (45)*	116	0.92	0.96	9.241e-01
multiple sentences (45)	65	0.67	0.67	4.252e-04
multiple sentences (45)	26	0.69	0.43	3.729e-03
multiple sentences (100)	128	0.73	0.98	2.050e-04
multiple sentences (100)	98	0.58	0.90	1.084e-03
multiple sentences (100)	52	0.41	0.59	5.536e-03

Table 6.1: *Comparison of mean statistics and Non-zero channels used between different models. The channel amount were in the case of CCA predetermined, while the sparse models train on all 128 channels. The filter length used for the filter model was 36 samples (144 ms). All statistics shown are how well the models explains the training data. No model was sufficient to generalize to evaluation data.*

7

Bibliography

Borga M. (1998). *Learning Multidimensional Signal Processing*. Ph.D. Thesis. Linköping University: Sweden.

Bronkhorst A. (2000). *The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multiple-Talker Conditions*. Acta Acustica united with Acustica, 86, pp. 117-128.

Bishop C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Gustafsson F., Ljung L., Millnert M. (2010). *Signal Processing*. Lund: Studentlitteratur.

Hastie T., Tibshirani R., Friedman J. (2008). *The Elements of Statistical Learning*. 5th print, Stanford.

Lalor E., Mesgarani N., Rajaram S., O'Donovan A., Wright J., Choi I., Brumberg J., Din N., Lee A. KC., Peters N., Ramenshalli S., Pompe J., Shinn-Cunningham B., Slaney M., Shamma S. (2012). *Decoding Auditory Attention (in Real Time) with EEG*. Neuromorphic Cognition Engineering Workshop.

Mesgarani N., Chang E. F. (2012). *Selective cortical representation of attended speaker in multi-talker speech perception*. Nature, 485, 233-236.

Meier L., Geer S., Bühlmann P. (2008). *The group lasso for logistic regression*. Royal Statistical Society, Vol 70, pp. 53-71.

Ng A., Jordan M. (2001). *On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes*. NIPS, pp 841-848. MIT Press.

Nordberg K. (2012). *Introduction to Homogeneous Representations and Estimation in Geometry*. Computer Vision Laboratory, Department of Electrical Engineering, Linköping University: Sweden

Podonik B., Stanley Eugene H. (2008). *Detrended Cross-Correlation Analysis: A New Method for Analyzing Two Non-stationary Time Series*. American Physical Society, Physical Review Letters, Vol 100, No 8.

Parra L., Spence C. (2000). *Convolutive Blind Separation of Non-Stationary Sources*. IEEE Transactions on Speech and Audio Processing, Vol. 8, No. 3.

Power A. J., Foxe J. J., Forde E., Reilly R. B., Lalor E. C. (2012). *At what time is the cocktail party? A late locus of selective attention to natural speech*. EJN, 35(1497-1503).

White L.B., Boashash B. (1990). *Cross Spectral Analysis of Non-Stationary Processes*. IEEE Transactions on Information Theory, Vol. 36, No. 4, pp. 830-835.

Zhan Y., Hilliday D., Jiang P., Liu X., Feng J. (2006). *Detecting time-dependent coherence between non-stationary electrophysiological signals–a combined statistical and time-frequency approach.*. Journal of neuroscience methods, Vol. 30, No. 156, pp. 322-32.

Viinikanoja J., Klami A., Kaski S. (2010). *Variational Bayesian Mixture of Robust CCA Models*. Lecture Notes in Computer Science, Vol 6323, pp 370-385, Springer: Berlin Heidelberg.

Appendix

A

Statistics

A statistic is a measure of some attribute of sampled data that are sampled from some distribution. Well known examples are the mean or the variance. The statistics used in this report are chosen with the intent that they should give a value of how similar two signals are to each other, but from different view points in order to capture a broader picture than if only using one of the solely would.

A.1 Correlation coefficient

The Pearson product-moment correlation coefficient is used within the context of this work for calculating the correlation coefficient between EEG signal (x) and sound envelope (y). If both have zero mean, the Pearson product-moment correlation coefficient is estimated by

$$\mathsf{cor}(x, y) = \frac{\sum_k x_c[k]y[k]}{\sqrt{\sum_k x_c^2[k] \sum_k y^2[k]}}, \tag{A.1}$$

where x_c denotes the channel c of the EEG signal and y the corresponding sound envelope.

A.2 Residuals

Let y be a signal and $\hat{y}$ be an estimate of y. The residual of the spectrograms $Y[m, f]$ and $\hat{Y}[m, f]$ is estimated by

$$\mathsf{Residual}(Y[m, f], \hat{Y}[m, f]) = \frac{1}{F-1} \frac{1}{M-1} \sum_f \sum_m (Y[m, f] - \hat{Y}[m, f])^2, \tag{A.2}$$

where m and f denote the temporal respectively frequency dimension, M and F denote the number of discrete times and frequencies respectively.

A.3 Explained Variance

Let y be a signal and $\hat{y}$ be an estimate of y. Explained variance give a value of how much of the variance in y that the model producing y' has captured. In general, the value will be between 0.0 and 1.0, however if the model introduces a higher variance than the signal the value will be less than 0. It is calculated as

$$\mathsf{EV}(y, \hat{y}) = 1 - \frac{Var(y - \hat{y})}{Var(y)}. \tag{A.3}$$

In the case of matrices and in our case of spectrograms it is estimated by

$$\mathsf{EV}(Y[m, f], \hat{Y}[m, f]) = 1 - \frac{\mathsf{Residual}(Y[m, f], \hat{Y}[m, f])}{\frac{1}{F-1}\frac{1}{M-1}\sum_f \sum_m (Y[m, f])^2}, \tag{A.4}$$

where m and f denote the temporal respectively frequency dimension, M and F denote the number of discrete times and frequencies respectively.